Fourier Self-Supervision for Fine-Grained Generalized Category Discovery

Sarah Rastegar[1], Mina Ghadimi Atigh[1], Pascal Mettes[1], Yuki M. Asano[2], and Cees G. M. Snoek[1]

[1] University of Amsterdam
[2] University of Technology Nuremberg

Abstract. Generalized Category Discovery aims to recognize known categories while identifying novel ones within unlabeled data. Existing methods, typically based on self-supervision and contrastive learning, often struggle to capture fine-grained distinctions, relying on superficial visual cues rather than the intrinsic attributes humans use for categorization. We introduce Fourier Self-Supervision, that leverages the Fourier transform of images to enhance the discrimination of subtle differences and support the discovery of new categories. Our method employs a dual frequency filtering strategy: a low-pass filter first extracts broad, abstract attributes that capture high-level category information, while a high-pass filter emphasizes fine details such as edges and textures that are essential for fine-grained recognition. Each operates on a dedicated latent space, and their overlapping representations together yield a richer, more complete feature space. This dual-frequency approach not only refines feature extraction to identify novel categories, but also strengthens the model's discriminative power in fine-grained category discovery. Experiments on multiple fine-grained datasets show that incorporating Fourier Self-Supervision outperforms state-of-the-art methods, even when the number of classes is unknown, demonstrating its effectiveness for Generalized Category Discovery. Our code is available at: https://github.com/SarahRastegar/FourEx.

1 Introduction

Generalized Category Discovery (GCD) mitigates the inability of supervised learning to recognize novel categories, by enabling the recognition of both known and novel categories within unlabeled datasets [1, 3, 11, 20, 43, 49, 50, 53, 62]. A prominent GCD strategy for addressing the challenge of unknown categories is self-supervision via contrastive learning [3,6,10,19,23,27,30,33,61], which utilizes data augmentation to produce different views of the same image [7,8,10,21,36]. Current literature mostly focuses on the coarse-grained setting, where categories have clear appearance differences. But as categories become more fine-grained and inter-class appearance differences start to shrink, we should still be able to figure out when a sample is part of a known set of labels, or is part of a new category. To tackle the

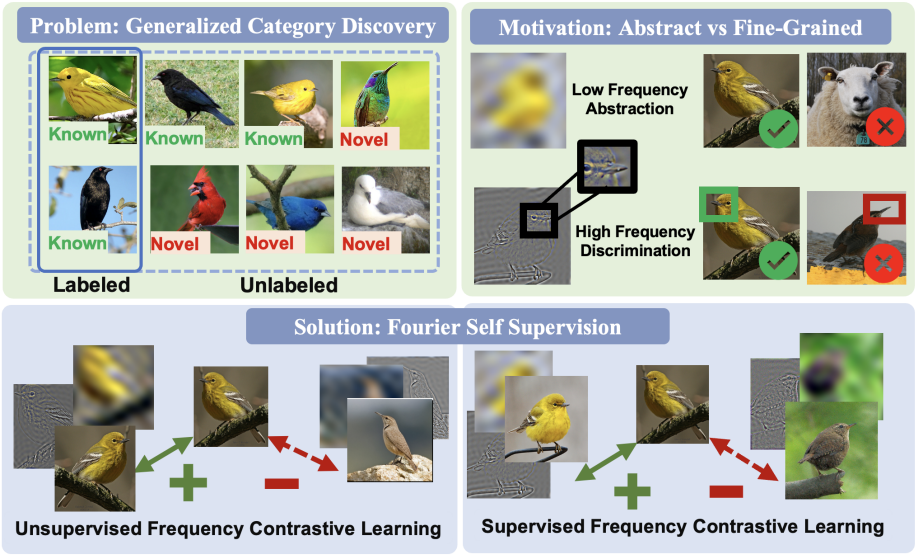

Fig. 1: Overview of Our Approach. *(a) Problem*, We aim to cluster unknown, novel categories alongside known ones. *(b) Motivation*, The blurred image of a yellow bird, reconstructed from low Fourier frequencies, provides a general sense of the category, facilitating broad generalizations. In contrast, high-frequency reconstructions reveal finer details, such as beaks and feathers. *(c) Fourier Self-Supervision* harnesses the generalizability of low frequencies and the fine-grained discrimination of high frequencies in Fourier self-supervision to discover novel, fine-grained categories.

fine-grained GCD challenge, we draw inspiration from Fourier analysis. Intuitively, an object's categorization should remain consistent across the frequency spectrum of its Fourier transform. However, broader, more general categories tend to manifest in lower frequencies, while fine-grained distinctions emerge more clearly in higher frequencies (see Fig. 2). Building on this, we introduce a frequency-based self-supervision to enhance category discovery, particularly when classes exhibit strong visual similarity. The overall problem setup, motivation, and a high-level overview of our method are illustrated in Fig. 1.

Incorporating Fourier self-supervision allows us to exploit the F-principle [57], which posits that overparameterized neural networks generalize well partly because they prioritize learning lower Fourier components of a function before gradually adjusting to higher frequencies. To extend generalizability to novel fine-grained categories, we aim for the model to infer unseen categories using the knowledge acquired from known ones. When categories adhere to an implicit hierarchical structure, observing a sibling category and its shared parent can help constrain the search space for novel categories. This implicit hierarchy, inferred through lower frequency components, allows the model to rapidly differentiate novel categories by focusing on their shared traits with known siblings, reducing both parent and sibling category confusion. For instance, in the left half of the

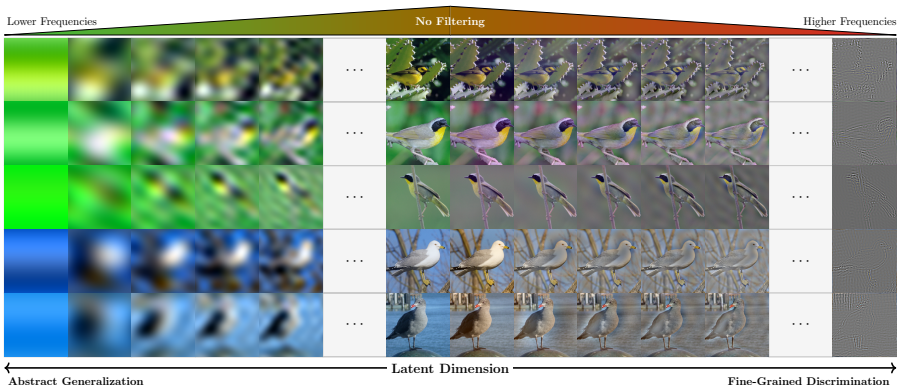

Fig. 2: Motivation for Fourier Self-Supervision. Low frequencies support broad category generalization (e.g., marine vs. forest birds), while high frequencies enable fine-grained recognition. By progressively integrating Fourier components and disentangling them in the latent space, our model builds hierarchical representations for both generalization and accuracy.

latent dimension in Fig. 2, lower frequencies provide a coarse-grained, general view of a category, allowing the model, for example, to differentiate whether a blurred image of a bird originates from a marine or forest environment. By progressively incorporating higher Fourier components, the model can more effectively distinguish fine-grained details as required for refined category discrimination. We propose an approach that assigns distinct parts of the latent space to capture these frequencies progressively, enabling the model to generalize to novel categories by learning an implicit hierarchy from low-frequency features, while high frequencies drive fine-grained differentiation.
Our contributions are as follows:

- We apply low-pass filtering in the Fourier domain to help the model capture broad, abstract characteristics of images, enabling it to infer an implicit hierarchical structure that improves generalization to unseen categories.
- We employ high-pass filtering to focus the model's attention on fine details and enhance its ability to distinguish subtle differences for fine-grained classification.
- Through empirical evaluation, we show that our method outperforms state-of-the-art approaches on fine-grained generalized category discovery tasks, while also achieving competitive performance on coarse-grained datasets.

2 Related Works

Generalized Category Discovery. The task of Generalized Category Discovery was formalized by Vaze *et al.* [49] and Cao *et al.* [3]. GCD lies at the intersection of supervised and unsupervised learning. It leverages a small amount

of labeled data along with a larger set of unlabeled data, where the unlabeled data can contain both known and novel categories. This makes GCD a special case of self-supervised learning [9, 35, 38, 45, 59]. There are two prominent approaches for GCD: (i) *Prototype-based Methods:* These approaches leverage a set of pre-defined prototypes as reference points to guide category discovery in the unlabeled data. This can be achieved through techniques like nearest neighbor search or learning a distance metric that effectively separates known and unknown categories, [1, 11, 20, 53]. (ii) *Similarity-driven Clustering:* This approach focuses on exploiting local similarities within the unlabeled data to form initial category clusters. This can be done through techniques like k-nearest neighbors or using mean-teacher frameworks to mitigate the issue of noisy pseudo-labels generated from similar information [2, 11, 20, 37, 42, 43, 50, 53, 62, 62]. Nonetheless, Contrastive learning methods often struggle with fine-grained GCD due to aggressive augmentations overshadowing subtle category differences [13]. Our work addresses this by leveraging the informative nature of high frequencies for fine-grained discrimination, while utilizing the lower frequencies for category discovery, enabling better handling of nuanced visual details. Several recent works explore alternative approaches for GCD. Hierarchical approaches proposed by Otholt *et al*. [37] and Banerjee *et al*. [2] leverage neighborhood structures for refined category delineation. Choi *et al*. [12] focus on robustness to noise by employing the mean shift algorithm for category discovery. Additionally, Wang *et al*. [52] propose a spatial prompt tuning method that incorporates spatial information from image data to focus better on specific object parts. Differing from all these works, our method leverages weak supervision by focusing on high frequencies to distinguish fine-grained categories from each other while focusing on low frequencies to generalize categorization to novel ones.

Fine Grained Generalized Category Discovery. One of the main challenges in generalized category discovery arises when categories are so fine-grained that traditional clustering in feature space struggles to accurately identify novel categories. To address this, prior work has taken several approaches: Fei *et al*. [16] partition data into expert sub-datasets by applying k-means clustering on self-supervised representations, while Rastegar *et al*. [43] proposed implicit category trees that enable hierarchical self-coding, preserving category similarity across multiple levels. Liu *et al*. [29] introduce class-wise distribution regularization to alleviate the bias towards known categories. Meanwhile, Rastegar *et al*. [44] introduced the concept of self-expertise, which leverages abstract pseudolabels to facilitate distinction between samples. Many recent methods [14, 22, 40, 58] exploit subtle sample-level variations to capture fine-grained details. In contrast, our approach selectively leverages frequency information, it emphasizes high-frequency components to enhance fine-grained discrimination, while relying on low frequencies to improve generalization to novel categories.

Fourier Transform based Image Augmentation. Fourier image augmentation has gained attention in computer vision due to its potential to enhance the robustness and generalizability of deep learning models. It leverages the Fourier transform to modify image data in the frequency domain, providing unique ad-

vantages over traditional spatial domain augmentations. Yang *et al*. [60] proposed Fourier Domain Adaptation, which applies the Fourier transform to both source and target domain images to align their frequency distributions. This improves classification accuracy considerably in domain adaptation scenarios by focusing on the low-frequency components that capture essential structural information while ignoring high-frequency noise. Sun *et al*. [46] introduced FourierMix, a data augmentation technique that blends images in the frequency domain. This method creates new training examples by mixing the amplitude and phase spectra of different images, leading to improved generalization performance on various benchmark datasets. Xu *et al*. [56] used Fourier magnitude swaps between different samples to improve domain generalization. In this work, we take inspiration from such methods and we introduce a self-supervised approach that leverages these properties of Fourier transforms to enable concept discovery in fine-grained settings, where details matter.

3 Preliminaries

Contrastive Learning for GCD. In a mini-batch with size B, let $\mathbf{x}_i$ and $\mathbf{x}'_i$ denote two distinct augmentations of the same image. The unsupervised contrastive loss $\mathcal{L}_i^u$ and its supervised counterpart $\mathcal{L}_i^s$ are formally described as [49]:

$$\mathcal{L}_i^u = -\log \frac{e^{\mathbf{z}_i . \mathbf{z}'_i/\tau}}{\sum_n \mathbb{1}_{[n\neq i]} e^{\mathbf{z}_i . \mathbf{z}'_n/\tau}}, \tag{1}$$

$$\mathcal{L}_i^s = -\frac{1}{|\mathcal{N}(i)|} \sum_{q\in\mathcal{N}(i)} \log \frac{e^{\mathbf{z}_i . \mathbf{z}_q/\tau}}{\sum_n \mathbb{1}_{[n\neq i]} e^{\mathbf{z}_i . \mathbf{z}_n/\tau}}. \tag{2}$$

In the proposed formulation, each image $\mathbf{x}_i$ is associated with an embedding $\mathbf{z}_i$, where $\mathbf{z}_i . \mathbf{z}'_i$ represents a similarity metric that may encompass cosine similarity, Euclidean distance, or other relevant measures. The indicator function $\mathbb{1}[n \neq i]$ is defined such that it equals 1 if and only if $n \neq i$, and τ is a temperature hyperparameter. Additionally, the unsupervised loss can be defined selectively for specific segments of the embedding vector. where $\mathcal{N}(i)$ denotes the collection of images within the same minibatch that are assigned the identical label as image $\mathbf{x}_i$. Finally, for generalized category discovery, Vaze *et al*. [49] consider the total loss as:

$$\mathcal{L}^g = (1-\lambda) \sum_{i\in B} \mathcal{L}_i^u + \lambda \sum_{i\in B_\mathcal{L}} \mathcal{L}_i^s. \tag{3}$$

where $B_\mathcal{L}$ is the labeled portion of minibatch B and λ is a mixing hyperparameter.
Fourier Transform. For an image $\mathbf{x}_i = f(h, y)$, its Fourier transformation $\mathcal{F}(u, v)$ and its reconstruction by using inverse transformation $\mathcal{F}^{-1}$is formulated as:

$$\mathcal{F}(u, v) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} f(h, w) e^{-2\pi i (\frac{h}{H}u + \frac{w}{W}v)}, \tag{4}$$

$$f(h,w)=\frac{1}{WH}\sum_{u=0}^{H-1}\sum_{v=0}^{W-1}\mathcal{F}(u,v)e^{2\pi i(\frac{u}{H}h+\frac{v}{W}w)}. \tag{5}$$

The efficient computation of both the Fourier transform and its inverse is facilitated by the Fast Fourier Transform (FFT) algorithm, as detailed by [34]. A low-pass filter (LPF), characterized by its transfer function $H_{LP}(u,v)$, incorporates a cutoff frequency denoted by f_l. This function defines the operational boundaries of the LPF as $H_{LP}(u,v)=\mathbb{1}(|u|<f_l,|v|<f_l)$. Correspondingly, a high pass filter (HPF) is formulated as $H_{HP}(u,v)=\mathbf{1}-H_{LP}(u,v)$ in which $\mathbf{1}$ represents the matrix of ones. The operation involves the convolution of the image with the respective filter kernel to implement either a low-pass or high-pass filter. This is efficiently achieved in the frequency domain by the element-wise multiplication of the image's Fourier transform with the filter's transfer function.

4 Fourier Self-Supervision

Inspired by the principles of Fourier optics [17] and recognizing that the human eye lens performs a natural Fourier transform, we propose that object classification can be discerned across multiple frequency bands. For this purpose, we define two specific cutoff frequencies, denoted as f_l and f_h. As shown in Fig. 3, to isolate lower frequencies, we employ a low-pass filter (LPF). Conversely, to eliminate lower frequency components, a high pass filter (HPF). In the subsequent sections, we provide a detailed description of how each filter is implemented.

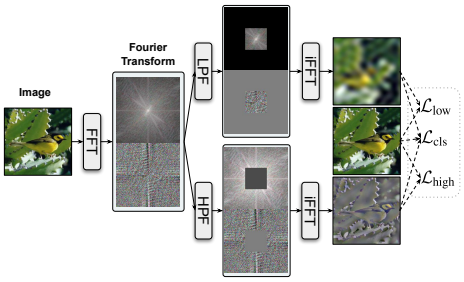

Fig. 3: Fourier Self-Supervision: An image undergoes a Fourier transform to extract its magnitude and phase. Both low-pass and high-pass filters are applied in parallel in the frequency domain. The inverse Fourier transform then reconstructs the frequency-filtered images, which are used as positive views for the original image in contrastive learning. Finally, all three views are utilized to predict the class of labeled images.

Extraction of Critical Frequencies. To effectively implement both low-pass and high-pass filtering, we define a threshold frequency to serve as an upper bound, ensuring that the model is not influenced by very high frequencies that typically contain noise, which can obscure informative patterns. We introduce a high threshold value, T_h, to cap these frequencies. For low-pass filtering, if the cutoff frequency is too high, the image retains excessive detail, offering little additional learning benefit beyond the original image. To make this threshold dataset-dependent rather than arbitrary, we determine the maximum threshold frequency that maintains a signal-to-noise ratio (SNR) of 15 dB, aligning with current standards for acceptable subjective image and video quality [41, 54]. To achieve category-based representations, we compute the high-frequency cutoff

threshold across all categories in the dataset and utilize the average value to ensure a balanced and dataset-dependent approach. The high-frequency cutoff is determined by maintaining a signal-to-noise ratio (SNR) that aligns with a standard of 15 dB, ensuring both informative features and noise suppression. To compute the SNR for different cutoff frequencies, we first represent the Fourier transform of an image as X, and calculate the SNR as $SNR(f_c)=10\log(\frac{\sum_{f\leq f_c}|X(f)|^2}{\sum_{f>f_c}|X(f)|^2})$, where $|.|^2$ is the average magnitude of a signal and f_c denotes the cutoff frequency. By averaging this threshold across all categories, we establish a robust, dataset-dependent criterion for frequency filtering, effectively balancing generalization and fine-grained discrimination. We then select an SNR of 32, equivalent to 15 dB, as this is widely recognized as the standard for acceptable signal quality. To ensure robustness, we also conduct experiments with different SNR values, as detailed in the appendix. Using this approach, we determine the low-pass cutoff frequency threshold T_l and its corresponding high-pass cutoff frequency threshold T_h. These thresholds are subsequently employed in the later stages of our method to guide both low-frequency generalization and high-frequency discrimination effectively.

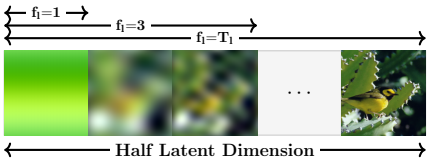

Fig. 4: Low-Frequency Contrastive Learning. Each image undergoes low-pass and high-pass filtering at a random cut-off frequency, denoted as f_l. Then the ratio $\frac{f_l}{T_l}$ is calculated, where T_l represents the maximum cutoff frequency. This ratio determines the fraction of the latent dimension dedicated to contrasting the filtered image against its original counterpart. In this example, a cutoff frequency of f_l=1, allocating only $\frac{1}{T_l}$ of its leftmost latent dimension for contrastive learning.

Low-Frequency Contrastive Learning. Human perception remains capable of discerning image categories despite the absence of high-frequency details, which are typically associated with noise or subtle information. As frequency components are progressively omitted, there is a loss of categorical information; for example, in Fig. 4, initial frequency bands allow only the recognition of basic attributes like color or the identification of an object as a bird. To leverage this abstracted information without compromising overall representational quality, we select a terminal frequency band T_l with an SNR of 15 dB or more, beyond which differences are imperceptible to the human eye. Utilizing $\frac{f_l}{T_l}$ of the dimensionality, we apply contrastive learning to facilitate generalized category discovery through embeddings. Considering an image $\mathbf{x}_i$, its low-frequency reconstruction which is denoted by $\mathbf{l}_i$ is $\mathbf{l}_i=\mathcal{F}^{-1}(H_{LP}\odot\mathcal{F}(\mathbf{x}_i))$. After the reconstruction phase, the image is employed as input for both supervised and unsupervised contrastive learning, specifically targeting a fraction $\frac{f_l}{T_l}$ of the latent dimension D. We define this fraction as $d=\frac{f_l}{T}D$ and incorporate it into Eq. (3) to compute the low-frequency contrastive loss:

$$\mathcal{L}_{\text{low}}=(1-\lambda)\sum_{i\in B}\mathcal{L}^u_{i|d}+\lambda\sum_{i\in B_{\mathcal{L}}}\mathcal{L}^s_{i|d}. \tag{6}$$

The loss $\mathcal{L}^u_{i|d}$ is defined based on the subset of its first d leftmost latent dimension.

High-Frequency Contrastive Learning. In this section, we explore high-frequency contrastive learning to enhance model sensitivity to fine-grained image details crucial for differentiating categories in fine-grained datasets. We select a random cutoff frequency $f_h{<}T_h$ to focus on these critical details. As depicted in Fig. 5, increasing f_h beyond a certain threshold introduces noise and reduces informativeness. To optimize the use of these nuanced details for category distinction without degrading the quality of representation, we employ a strategy analogous to low-frequency contrastive learning. Here, we initiate from the rightmost latent representation, unlike the leftmost starting point used in low-frequency settings. Consider a frequency band T_h, beyond which category distinction becomes imperceptible to humans. We allocate the last $1{-}\frac{f_h}{T_h}$ of the dimension for contrastive learning. Considering an image $\mathbf{x}_i$, its high-frequency reconstruction is denoted by $\mathbf{h}_i$ which can be calculated as $\mathbf{h}_i{=}\mathcal{F}^{-1}(H_{HP}{\odot}\mathcal{F}(\mathbf{x}_i))$ Similar to the previous section, we integrate $\mathbf{h}_i$ as an alternative view of $\mathbf{x}_i$. For contrastive learning, we employ the last $d{=}\frac{f_h}{T}D$ dimensions, denoted as $-d$. This substitution is utilized in Eq. (3) to compute the high-frequency contrastive loss.

$$\mathcal{L}_{\text{high}}{=}(1{-}\lambda)\sum_{i\in B}\mathcal{L}^u_{i|-d}{+}\lambda\sum_{i\in B_{\mathcal{L}}}\mathcal{L}^s_{i|-d}. \tag{7}$$

In the Appendix, we show that fine details reside only in high-frequency modes; dropping them discards the information needed to distinguish those details. In practice, rather than allocating the entire D latent dimension to both low and high-frequency components, we divide it such that the left half is dedicated to the low-frequencies, while the right half is assigned to the high-frequencies.

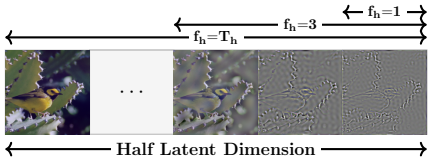

Fig. 5: High-Frequency Contrastive Learning. Each image undergoes high-pass filtering at a random cutoff frequency, denoted as f_h. Then the ratio $\frac{f_h}{T_h}$ is calculated, where T_h represents the maximum cutoff frequency. This ratio determines the fraction of the latent dimension dedicated to contrasting the filtered image against its original counterpart. In this example, the square corresponding to 1 is filtered at a cutoff frequency of $T_h - 1$, allocating only $\frac{1}{T_h}$ of its rightmost latent dimension for contrastive learning.

Frequency Augmentation Classification. In addition to employing contrastive learning, our approach also acknowledges that the low-frequency reconstruction $\mathbf{l}_i$, high-frequency reconstruction $\mathbf{h}_i$, and the original sample $\mathbf{x}_i$ share identical ground truth labels. Consequently, we incorporate a binary cross-entropy loss function to facilitate the label classification from these instances. This enables a unified treatment of label consistency across different representation modalities within our framework.

$$\mathcal{L}_{\text{cls}}{=}\mathcal{L}_{\text{BCE}}(p_{h_i}, y_i){+}\mathcal{L}_{\text{BCE}}(p_{l_i}, y_i){+}\mathcal{L}_{\text{BCE}}(p_{x_i}, y_i). \tag{8}$$

In this formulation, p_{h_i} denotes the labels predicted by the model for $\mathbf{h}_i$. Summing up, the total loss is computed as the aggregation of individual losses together with the baseline model's loss, thus we have:

$$\mathcal{L}_{\text{total}}=\mathcal{L}_{\text{base}} + \alpha_{\text{low}}\mathcal{L}_{\text{low}} + \alpha_{\text{high}}\mathcal{L}_{\text{high}} + \alpha_{\text{cls}}\mathcal{L}_{\text{cls}}. \tag{9}$$

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate our method on three fine-grained datasets: CUB-200 [51], FGVC-Aircraft [32], and Stanford-Cars [24]. Further, we include the challenging Herbarium dataset [47], which is both fine-grained and long-tailed, demonstrating that our approach remains effective even under severe class imbalance. We also report results on Oxford-IIIT Pet [39], a small-scale fine-grained dataset with limited samples per category, making it particularly challenging. In the Appendix, we further evaluate our approach on coarse-grained datasets, CIFAR-10 [25], CIFAR-100 [25], and ImageNet-100 [15] to demonstrate its adaptability to both fine-grained and coarse-grained classification. Detailed dataset statistics and train/test split information are also included in the Appendix.

Implementation Details. In our experiments, we adopted the data partitioning strategy by Vaze *et al.* [49], wherein half the categories for each dataset are recognized as known. From these known categories, half of the samples form the labeled set, while the residual data from known categories, along with all samples from novel categories, are allocated to the unlabeled set. Following the approach by Vaze *et al.* [49], our primary model architecture is the ViT-B/16, utilizing either a DINOv1 [8] pre-training on the unlabeled ImageNet1K [26] dataset or a DINOv2 [36] pre-training on the LVD-142M dataset. We apply Fourier Self-Supervision to baselines SimGCD [53] and SelEx [44], resulting in two variants we call *FourSim* and *FourEx*, respectively. For FourSim, we froze the initial 11 blocks of the ViT-B/16 and fine-tuned the last block, while for FourEx, we finetuned last three blocks. In the Appendix, we show that independent of the number of frozen blocks, our method improves over both baselines. Implementation details for augmenting other backbones with Fourier Self-Supervision are provided in the Appendix.

Before training, we measure the signal-to-noise ratio (SNR) across a range of cutoff frequencies and select the cutoff that produces an SNR of 15 dB, a standard commonly used in video quality assessment. This procedure only requires applying the Fourier transform to a batch of training images to estimate the SNR for different cutoff frequencies. The corresponding cutoff values for different datasets and SNR levels are provided in the Appendix. In practice, we observe that the granularity of a dataset largely determines the resulting SNR values. Furthermore, to mitigate Gibbs ringing artifacts near image edges, we apply a Gaussian blur with a kernel size of 5 to each image before filtering.

Table 1: Comparison with state-of-the-art for fine-grained image classification. Bold and underlined numbers indicate the best and second-best accuracies, respectively. Our method is well-suited for fine-grained datasets, profits from stronger backbones, and has strong performance for all three experimental settings (*All*, *Known*, and *Novel*).

			CUB-200			FGVC-Aircraft			Stanford-Cars			Average		
	Method	**Venue**	All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel
DINOv1	GCD [49]	*CVPR22*	51.3	56.6	48.7	45.0	41.1	46.9	39.0	57.6	29.9	45.1	51.8	41.8
	PromptCAL [62]	*CVPR23*	62.9	64.4	62.1	52.2	52.2	52.3	50.2	70.1	40.6	55.1	62.2	51.7
	μGCD [50]	*NeurIPS23*	65.7	68.0	64.6	53.8	55.4	53.0	56.5	68.1	50.9	58.7	63.8	56.2
	InfoSieve [43]	*NeurIPS23*	69.4	**77.9**	65.2	56.3	63.7	52.5	55.7	74.8	46.4	60.5	72.1	54.7
	SPTNet [52]	*ICLR24*	65.8	68.8	65.1	59.3	61.8	58.1	59.0	79.2	49.3	61.4	69.9	57.5
	CMS [12]	*CVPR24*	68.2	76.5	64.0	56.0	63.4	52.3	56.9	76.1	47.6	60.4	72.0	54.6
	TRAILER [55]	*CVPR24*	65.1	71.3	54.5	61.9	62.6	50.5	55.4	71.7	47.6	61.0	64.5	54.7
	LegoGCD [5]	*CVPR24*	63.8	71.9	59.8	55.0	61.5	51.7	57.3	75.7	48.4	58.7	69.7	53.3
	SelEx [44]	*ECCV24*	73.6	75.3	72.8	57.1	64.7	53.3	58.5	75.6	50.3	63.0	71.9	58.8
	FlipClass [28]	*NeurIPS24*	71.3	71.3	71.3	59.3	66.9	55.4	63.1	**81.7**	53.8	64.6	73.3	60.2
	DebGCD [31]	*ICLR25*	66.3	71.8	63.5	61.7	63.9	60.6	**65.3**	81.6	57.4	64.4	72.4	60.5
	APL [14]	*CVPR25*	68.5	73.1	66.2	60.9	63.5	59.6	62.3	80.7	53.4	63.9	72.4	59.7
	MOS [40]	*CVPR25*	69.6	72.3	68.2	61.1	66.9	58.2	64.6	80.9	56.7	65.1	73.4	61.0
	AllGCD [4]	*ICCV25*	68.4	75.1	65.1	57.4	62.4	54.9	60.5	77.0	52.5	62.1	71.5	57.5
	SEAL [22]	*NeurIPS25*	66.2	72.1	63.2	62.0	65.3	60.4	**65.3**	79.3	**58.5**	64.5	72.2	60.7
	SimGCD [53]	*ICCV23*	60.3	65.6	57.7	54.2	59.1	51.8	53.8	71.9	45.0	56.1	65.5	51.5
	FourSim (Ours)		70.3	75.5	67.7	56.1	63.1	52.6	56.6	78.9	45.8	61.0	72.5	55.4
	SelEx† [44]	*ECCV24*	76.4	72.4	78.4	61.2	67.7	58.0	56.9	76.9	47.3	64.8	72.3	61.2
	FourEx (Ours)		**80.2**	75.1	**82.7**	**65.9**	**67.9**	**64.9**	59.2	76.9	50.6	**68.4**	73.3	**66.1**
	Avg Δ		+6.9	+6.3	+7.2	+3.3	+2.1	+3.9	+2.6	+3.5	+2.1	+4.3	+4.0	+4.4
DINOv2	GCD* [49]	*CVPR22*	71.9	71.2	72.3	55.4	47.9	59.2	65.7	67.8	64.7	64.3	62.3	65.4
	SimGCD* [53]	*ICCV23*	71.5	78.1	68.3	63.9	69.9	60.9	71.5	81.9	66.6	69.0	76.6	65.3
	μGCD* [50]	*NeurIPS23*	74.0	75.9	73.1	66.3	68.7	65.1	76.1	91.0	68.9	72.1	78.5	69.0
	FlipClass [28]	*NeurIPS24*	79.3	80.7	78.5	71.1	75.1	69.1	78.0	88.0	73.2	76.1	81.3	73.6
	DebGCD [31]	*ICLR25*	77.5	80.8	75.8	71.9	76.0	69.8	75.4	87.7	69.5	74.9	81.5	71.7
	SEAL [22]	*NeurIPS25*	76.7	78.3	75.9	74.6	73.2	75.3	77.7	88.7	72.4	76.3	80.1	74.5
	SelEx [44]	*ECCV24*	87.4	85.1	88.5	79.8	82.3	78.6	**82.2**	93.7	**76.7**	**83.1**	87.0	**81.3**
	FourEx (Ours)		**87.8**	**86.3**	**88.6**	**81.5**	**84.7**	**80.0**	80.1	**94.3**	73.2	**83.1**	**88.4**	80.6
	Avg Δ		+0.4	+1.2	+0.1	+1.7	+2.4	+1.4	-2.1	+0.6	-3.5	+0.0	+1.4	-0.7

* Reported from [50]. † Results for fine-tuning the last three blocks.

5.2 Comparison with State-of-the-Art

Fine-grained image classification. In Tab. 1, we evaluate the performance of FourEx on three fine-grained datasets. The results show that our method excels at fine-grained categorization, outperforming existing approaches in both the *All* and *Novel* category settings for both DINOv1 and DINOv2 backbones. This improvement arises from the model's ability to separate subtle high-frequency details from broader low-frequency patterns, enabling more precise recognition of fine-grained categories. Both *FourSim* and *FourEx*, consistently improve over their respective baselines when using a DINOv1 backbone. A similar trend is observed with DINOv2, with the exception of the Stanford Cars dataset. This dataset is comparatively less fine-grained than the others, as the differences between cars are more pronounced. Consequently, the improvements of our method are more evident on finer-grained datasets.

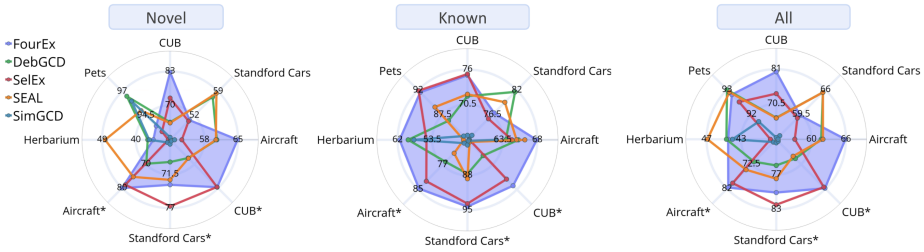

Fig. 6: Radar chart comparing our proposed method against state-of-the-art approaches across five fine-grained datasets: CUB, Stanford Cars, Aircraft, Oxford Pets, and the long-tailed Herbarium. Datasets marked with * utilize the DINOv2 backbone. Our method visually envelops the baselines on Known and All metrics, while consistently outperforming the baseline, SelEx.

As shown in Tab. 2, our model also performs consistently well on the Oxford-IIIT Pets dataset across all evaluation settings. Although this dataset is small and prone to overfitting, Fourier-based augmentation effectively expands the data distribution, mitigating overfitting risks. Moreover, our method demonstrates noticeable improvement over the baseline on long-tailed fine-grained datasets such as Herbarium, maintaining competitive performance despite its class imbalance.

Fig. 6 illustrates the comparison between our method and state-of-the-art approaches across multiple datasets. As shown, while methodslike SEAL [22] achieve top performance on Stanford Cars and Herbarium, the radar chart reveals that these gains on novel categories come at the expense of accuracy on known categories. Similarly, although DebGCD [31] leads on novel categories for Oxford Pets and known categories for Stanford Cars, our method consistently outperforms it in overall aggregate metrics. This demonstrates that our approach offers a more robust balance across diverse datasets with different granularities.

Table 2: Comparison with state-of-the-art GCD methods on Herbarium19 [47] and Oxford-Pet [39] on DINOv1. Bold and underlined numbers indicate the best and second-best accuracies, respectively. Our method is well suited for fine-grained datasets, profits from stronger backbones, and has strong performance for all three experimental settings.

	Oxford-Pet			Herbarium19		
Method	All	Known	Novel	All	Known	Novel
GCD [49]	80.2	85.1	77.6	35.4	51.0	27.0
SimGCD [53]	91.7	83.6	<u>96.0</u>	44.0	58.0	36.4
InfoSieve [43]	91.8	**92.6**	91.3	40.3	59.0	30.2
DebGCD [31]	**93.0**	86.4	**96.5**	<u>44.7</u>	<u>59.4</u>	<u>36.8</u>
SEAL [22]	<u>92.9</u>	88.9	95.0	**46.9**	45.8	**48.2**
SelEx [44]	92.5	<u>91.9</u>	92.8	39.6	54.9	31.3
FourEx (Ours)	92.8	91.6	93.4	44.5	**61.5**	35.3
Avg Δ	+0.3	-0.3	+0.6	+4.9	+6.6	+4.0

Estimate the number of categories. In our earlier experiments, we assumed prior knowledge of the exact number of categories, a condition that may not hold in practical applications. To evaluate the robustness of FourEx under more realistic scenarios, we tested its performance when the number of categories is estimated rather than known. We employed the estimation technique proposed by Vaze *et al.* [49], which yielded

an estimated count of 231 categories for the CUB and 230 for Stanford-Cars dataset. As shown in Tab. 3, FourEx outperforms existing methods across almost all evaluation metrics (*All*, *Known*, *Novel*), demonstrating its effectiveness even when the category count is uncertain.

Coarse-grained image classification. Finally, as detailed in the Appendix, while primarily designed for fine-grained scenarios, our approach also achieves competitive results on coarse-grained datasets, highlighting its versatility.

Table 3: Comparison with state-of-the-art with Estimated Number of Categories on DINOv1 Bold numbers show the best accuracies. We leverage the estimation technique from Vaze *et al.* [49] Our approach has competitive performance with other methods in all scenarios.

	CUB-200			Stanford-Cars		
Method	All	Known	Novel	All	Known	Novel
GCD [49]	47.1	55.1	44.8	39.1	58.6	29.7
SimGCD [53]	61.5	66.4	59.1	49.1	65.1	41.3
GCA [37]	62.0	65.2	60.4	54.4	72.1	45.8
μGCD [50]	62.0	60.3	62.8	56.3	66.8	**51.1**
PIM [48]	62.0	75.7	55.1	42.4	65.3	31.3
CMS [12]	64.4	68.2	62.2	51.7	68.9	43.4
SelEx [44]	72.0	72.3	71.9	58.7	75.3	50.8
Yang *et. al.* [58]	61.3	60.8	62.1	44.3	58.2	39.1
NC-GCD [18]	70.3	72.1	69.4	54.0	73.1	44.8
FourEx (Ours)	**77.3**	**75.9**	**77.9**	**58.8**	**77.6**	49.7

5.3 Ablative studies

In this section, we analyze the individual contributions of our method's components and hyperparameters. All experiments are conducted on CUB-200 using the DINOv1 backbone. Additional ablations with other datasets, backbones, baselines, and qualitative results are provided in the Appendix.

Table 4: Effectiveness of model components. We evaluate the contribution of each model component across three fine-grained datasets. All modules contribute to improved performance on both known and novel categories. Note that the baseline is SelEx with the last three blocks fine-tuned.

Component			CUB-200			FGVC-Aircraft			Standford-Cars			Avg
$\mathcal{L}_{\text{low}}$	$\mathcal{L}_{\text{high}}$	$\mathcal{L}_{\text{cls}}$	**All**	**Known**	**Novel**	**All**	**Known**	**Novel**	**All**	**Known**	**Novel**	**All**
			76.4	72.4	78.4	61.2	67.7	58.0	56.9	76.9	47.3	64.8
✓			77.5	74.0	79.3	63.8	67.7	61.9	56.3	77.1	46.3	65.9
	✓		78.8	72.5	81.9	63.9	63.6	64.1	56.8	74.0	48.5	66.5
		✓	78.0	71.2	81.4	62.8	65.5	61.4	57.1	76.3	47.8	66.0
✓	✓		80.0	75.7	82.1	64.9	69.6	62.5	57.7	78.2	47.8	67.5
✓		✓	77.9	75.3	79.2	64.5	66.3	63.6	57.7	77.7	48.0	66.7
	✓	✓	79.7	72.8	83.1	63.2	66.1	61.7	57.9	76.2	49.1	66.9
✓	✓	✓	80.2	75.1	82.7	65.9	67.9	64.9	59.2	76.9	50.6	68.4

Effect of each component. Tab. 4 examines the effect of our three key method components: low-frequency contrastive learning ($\mathcal{L}_{\text{low}}$), high-frequency contrastive learning ($\mathcal{L}_{\text{high}}$), and frequency augmentation classification ($\mathcal{L}_{\text{cls}}$). The results demonstrate distinct benefits of low-frequency and high-frequency contrastive learning for novel and known categories, respectively. Our low-frequency contrastive learning, denoted as $\mathcal{L}_{\text{low}}$, shows a particular affinity for enhancing both known and novel categories. This aligns with our initial hypothesis that generalization for both set of categories benefit from more abstraction. Low-frequency

signals create an implicit hierarchy in the training samples, necessitating correct categorization based on broad context. Conversely, our high-frequency contrastive learning, denoted as $\mathcal{L}_{\text{high}}$, excels in aiding novel categories. This is because hierarchical structures are advantageous for novel categories, but fine-grained details require the nuanced distinctions provided by high-frequency information. Interestingly, our classification objective alone tends to degrade performance for known categories. However, when decoupled with our contrastive losses, it improves results on novel categories. We attribute this to the prevention of overfitting, where the combined losses provide a more balanced training regimen.

Table 5: Hyperparameter analysis. These tables show the impact of each hyperparameter on the CUB dataset for both known and novel categories: (a) low-frequency coefficient α_{low} (b) high-frequency coefficient α_{high} (c) classification coefficient α_{cls} and (d) the SNR for the cutoff frequency.

(a) Effect of Low Frequency Hyperparameter

Low Frequency Coef	All	Known	Novel
$\alpha_{\text{low}} = 0$	77.9	72.1	80.7
$\alpha_{\text{low}} = 0.01$	80.2	74.7	83.0
$\alpha_{\text{low}} = 0.1$	80.2	75.1	82.7
$\alpha_{\text{low}} = 0.5$	76.8	74.3	78.0
$\alpha_{\text{low}} = 1$	69.2	70.1	68.8

(b) Effect of High Frequency Hyperparameter

High Frequency Coef	All	Known	Novel
$\alpha_{\text{high}} = 0$	78.1	76.3	78.9
$\alpha_{\text{high}} = 0.01$	78.3	75.8	79.6
$\alpha_{\text{high}} = 0.1$	80.2	75.1	82.7
$\alpha_{\text{high}} = 0.5$	79.2	75.0	81.3
$\alpha_{\text{high}} = 1$	77.1	73.7	78.7

(c) Effect of Classification Hyperparameter

Classification Coef	All	Known	Novel
$\alpha_{\text{cls}} = 0$	79.7	76.5	81.4
$\alpha_{\text{cls}} = 0.01$	78.9	73.1	81.8
$\alpha_{\text{cls}} = 0.1$	80.2	75.1	82.7
$\alpha_{\text{cls}} = 0.5$	79.6	73.3	82.7
$\alpha_{\text{cls}} = 1$	73.1	74.9	72.3

(d) Effect of SNR of Cuttoff Frequency

SNR Threshold Choice	All	Known	Novel
SNR=5 dB	79.0	75.7	80.7
SNR=10 dB	79.2	72.3	82.6
SNR=15 dB	80.2	75.1	82.7
SNR=20 dB	78.4	74.3	80.5
SNR=25 dB	77.1	74.3	78.5

Effects of α_{low}. This hyperparameter represents the coefficient for the low-frequency component. As shown in Tab. 5 (a) increasing α_{low} initially improves the model's performance on both novel and known categories. However, beyond a certain threshold, further increasing this hyperparameter begins to degrade performance. This behavior is intuitive, as the model relies on a blurred version of the data for this loss; excessive emphasis on low-frequency information causes the model to lose critical fine-grained details necessary for accurate categorization. As expected, the positive impact of this hyperparameter is more pronounced on novel categories, where the generalization facilitated by low-frequency components plays a larger role.

Effects of α_{high}. This hyperparameter represents the coefficient for high-frequency components. As shown in Tab. 5 (b), increasing α_{high} improves the model's performance on novel categories, while this hyperparameter negatively impacts known categories performance. This behavior aligns with the fact that high frequencies inherently carry lower energy compared to the overall image.

Fig. 7: Failure Cases of Our Model. Many errors arise from occlusion, reflection, or color mismatches. While the model misclassifies in these examples, it often identifies a related species within the same hierarchy.

When the model overemphasizes these low-energy components, it struggles to maintain stability when integrating the higher-energy low-frequency components, leading to degraded performance.

Effects of α_{cls}. In our experiments, α_{cls} represents the classification coefficient applied to both low-frequency and high-frequency constructed images. As shown in Tab. 5 (c), increasing this hyperparameter improves performance for novel categories. However, this trend diminishes beyond a certain threshold for α_{cls}, indicating that the model might overfit to the low/high-frequency reconstructions.

Effects of Different SNRs. In our experiments, we primarily adopt an SNR of 15 dB, as it represents the standard minimum threshold for acceptable subjective quality in most practical scenarios. To further assess the impact of signal quality on model performance, we evaluate the model across a range of SNR values. The results, summarized in Tab. 5 (d), demonstrate the model's adaptability to varying SNR levels. Notably, lower SNRs tend to favor generalization to novel categories, while higher SNRs improve adaptation to known categories. However, we observe that beyond 20 dB, the performance gains plateau, indicating that increasing resolution beyond this threshold does not significantly enhance the model's ability to distinguish categories. This suggests a practical limit to the benefits of resolution in generalized category discovery.

Failure Cases. Figure 7, illustrates instances where our model misclassified the data. A notable number of these errors are attributed to reflection or occlusion. Additionally, color appears to be a contributing factor. Since our model's high-pass filter removes color information, it becomes more susceptible to these types of misclassifications.

6 Conclusion

In this paper, we introduce a frequency-based augmentation strategy for self-supervised learning aimed at improving fine-grained generalized category discovery. By leveraging the complementary roles of low- and high-frequency information, our approach enables models to capture both broad, abstract patterns that support generalization and subtle details essential for precise discrimination. Low-frequency contrastive learning encourages the emergence of implicit category hierarchies, while high-frequency learning enhances sensitivity to fine-grained distinctions. Our method consistently improves performance on fine-grained benchmarks and achieves competitive results on long-tailed and coarse-grained datasets (see Appendix), demonstrating robustness across varying levels of category granularity. Moreover, the approach is plug-and-play and can be readily integrated into existing architectures with minimal additional computational overhead, making it practical and widely applicable.

Acknowledgments

This work is part of the project Real-Time Video Surveillance Search with project number 18038, which is (partly) financed by the Dutch Research Council (NWO) domain Applied and Engineering/Sciences (TTW).

References

1. An, W., Tian, F., Zheng, Q., Ding, W., Wang, Q., Chen, P.: Generalized category discovery with decoupled prototypical network. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37 (2023) 1, 4
2. Banerjee, A., Kallooriyakath, L.S., Biswas, S.: Amend: Adaptive margin and expanded neighborhood for efficient generalized category discovery. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2101–2110 (2024) 4
3. Cao, K., Brbic, M., Leskovec, J.: Open-world semi-supervised learning. In: Proceedings of the International Conference on Learning Representations (2022) 1, 3
4. Cao, X., Chen, K., Yang, F., Zheng, X., Tian, Y., Lu, Y.: Allgcd: Leveraging all unlabeled data for generalized category discovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3293–3303 (2025) 10
5. Cao, X., Zheng, X., Wang, G., Yu, W., Shen, Y., Li, K., Lu, Y., Tian, Y.: Solving the catastrophic forgetting problem in generalized category discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16880–16889 (2024) 10
6. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European conference on computer vision (ECCV). pp. 132–149 (2018) 1
7. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020) 1

8. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021) 1, 9
9. Chapelle, O., Scholkopf, B., Zien, A.: Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. IEEE Transactions on Neural Networks **20**(3), 542–542 (2009) 4
10. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020) 1
11. Chiaroni, F., Dolz, J., Masud, Z.I., Mitiche, A., Ben Ayed, I.: Parametric information maximization for generalized category discovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1729–1739 (2023) 1, 4
12. Choi, S., Kang, D., Cho, M.: Contrastive mean-shift learning for generalized category discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 4, 10, 12
13. Cole, E., Yang, X., Wilber, K., Mac Aodha, O., Belongie, S.: When does contrastive visual representation learning work? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14755–14764 (2022) 4
14. Dai, Q., Huang, H., Wu, Y., Yang, S.: Adaptive part learning for fine-grained generalized category discovery: A plug-and-play enhancement. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25444–25453 (2025) 4, 10
15. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009) 9
16. Fei, Y., Zhao, Z., Yang, S., Zhao, B.: Xcon: Learning with experts for fine-grained category discovery. In: British Machine Vision Conference (2022) 4
17. Goodman, J.W.: Introduction to Fourier optics. Roberts and Company publishers (2005) 6
18. Han, J., Wang, S., He, Y., Ding, C., Wang, Q., Gao, X., Dong, S., Gong, Y.: Consistent supervised-unsupervised alignment for generalized category discovery. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems 12
19. Han, K., Rebuffi, S.A., Ehrhardt, S., Vedaldi, A., Zisserman, A.: Automatically discovering and learning new visual categories with ranking statistics. In: Proceedings of the International Conference on Learning Representations (2020) 1
20. Hao, S., Han, K., Wong, K.Y.K.: CiPR: An efficient framework with cross-instance positive relations for generalized category discovery. Transactions on Machine Learning Research (2024) 1, 4
21. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020) 1
22. He, Z., Liu, Y., Han, K.: SEAL: Semantic-aware hierarchical learning for generalized category discovery. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025) 4, 10, 11
23. Jaiswal, A., Babu, A.R., Zadeh, M.Z., Banerjee, D., Makedon, F.: A survey on contrastive self-supervised learning. Technologies **9**(1), 2 (2020) 1
24. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 554–561 (2013) 9

25. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. Rep. 0, University of Toronto, Toronto, Ontario (2009) 9
26. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 25. Curran Associates, Inc. (2012) 9
27. Li, J., Zhou, P., Xiong, C., Hoi, S.: Prototypical contrastive learning of unsupervised representations. In: International Conference on Learning Representations (2021) 1
28. Lin, H., An, W., Wang, J., Chen, Y., Tian, F., Wang, M., Dai, G., Wang, Q., Wang, J.: Flipped classroom: Aligning teacher attention with student in generalized category discovery. In: Thirty-eighth Conference on Neural Information Processing Systems (2024) 10
29. Liu, D., Tan, Z., Zhao, L., Zhang, Z., Fang, X., Huang, W.: Generalized category discovery via reciprocal learning and class-wise distribution regularization. arXiv preprint arXiv:2506.02334 (2025) 4
30. Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J.: Self-supervised learning: Generative or contrastive. IEEE Transactions on Knowledge and Data Engineering **35**(1), 857–876 (2021) 1
31. Liu, Y., Han, K.: DebGCD: Debiased learning with distribution guidance for generalized category discovery. In: The Thirteenth International Conference on Learning Representations (2025) 10, 11
32. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013) 9
33. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: European Conference on Computer Vision. pp. 69–84. Springer (2016) 1
34. Nussbaumer, H.J., Nussbaumer, H.J.: The fast Fourier transform. Springer (1982) 6
35. Oliver, A., Odena, A., Raffel, C.A., Cubuk, E.D., Goodfellow, I.: Realistic evaluation of deep semi-supervised learning algorithms. In: Advances in neural information processing systems. vol. 31 (2018) 4
36. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024) 1, 9
37. Otholt, J., Meinel, C., Yang, H.: Guided cluster aggregation: A hierarchical approach to generalized category discovery. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2618–2627 (2024) 4, 12
38. Ouali, Y., Hudelot, C., Tami, M.: An overview of deep semi-supervised learning. arXiv preprint arXiv:2006.05278 (2020) 4
39. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3498–3505. IEEE (2012) 9, 11
40. Peng, Z., Ma, J., Sun, Z., Yi, R., Song, H., Tan, X., Ma, L.: Mos: Modeling object-scene associations in generalized category discovery. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15118–15128 (2025) 4, 10
41. Poynton, C.: Digital Video and HD: Algorithms and Interfaces. Morgan Kaufmann Publishers (2012) 6

42. Pu, N., Zhong, Z., Sebe, N.: Dynamic conceptional contrastive learning for generalized category discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 4
43. Rastegar, S., Doughty, H., Snoek, C.: Learn to categorize or categorize to learn? self-coding for generalized category discovery. In: Thirty-seventh Conference on Neural Information Processing Systems (2023) 1, 4, 10, 11
44. Rastegar, S., Salehi, M., Asano, Y.M., Doughty, H., Snoek, C.G.M.: Selex: Self-expertise in fine-grained generalized category discovery. In: ECCV (2024) 4, 9, 10, 11, 12
45. Rebuffi, S.A., Ehrhardt, S., Han, K., Vedaldi, A., Zisserman, A.: Semi-supervised learning with scarce annotations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 762–763 (2020) 4
46. Sun, J., Mehra, A., Kailkhura, B., Chen, P.Y., Hendrycks, D., Hamm, J., Mao, Z.M.: A spectral view of randomized smoothing under common corruptions: Benchmarking and improving certified robustness. In: European Conference on Computer Vision. pp. 654–671. Springer (2022) 5
47. Tan, K.C., Liu, Y., Ambrose, B., Tulig, M., Belongie, S.: The herbarium challenge 2019 dataset. arXiv preprint arXiv:1906.05372 (2019) 9, 11
48. Tan, Z., Yang, X., Wang, Q., Nguyen, A., Huang, K.: Interpret your decision: Logical reasoning regularization for generalization in visual classification. In: Thirty-eighth Conference on Neural Information Processing Systems (2024) 12
49. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Generalized category discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022) 1, 3, 5, 9, 10, 11, 12
50. Vaze, S., Vedaldi, A., Zisserman, A.: No representation rules them all in category discovery. Advances in Neural Information Processing Systems 37 (2023) 1, 4, 10, 12
51. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset (Jul 2011) 9
52. Wang, H., Vaze, S., Han, K.: SPTNet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. In: The Twelfth International Conference on Learning Representations (2024) 4, 10
53. Wen, X., Zhao, B., Qi, X.: Parametric classification for generalized category discovery: A baseline study. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16590–16600 (2023) 1, 4, 9, 10, 11, 12
54. Winkler, S.: Digital video quality: vision models and metrics. John Wiley & Sons (2005) 6
55. Xiao, R., Feng, L., Tang, K., Zhao, J., Li, Y., Chen, G., Wang, H.: Targeted representation alignment for open-world semi-supervised learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23072–23082 (2024) 10
56. Xu, Q., Zhang, R., Zhang, Y., Wang, Y., Tian, Q.: A fourier-based framework for domain generalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14383–14392 (2021) 5
57. Xu, Z.Q.J., Zhang, Y., Xiao, Y.: Training behavior of deep neural network in frequency domain. In: Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part I 26. pp. 264–274. Springer (2019) 2
58. Yang, F., Pu, N., Li, W., Luo, Z., Li, S., Sebe, N., Zhong, Z.: Learning to distinguish samples for generalized category discovery. In: European Conference on Computer Vision. pp. 105–122. Springer (2024) 4, 12

59. Yang, X., Song, Z., King, I., Xu, Z.: A survey on deep semi-supervised learning. IEEE Transactions on Knowledge and Data Engineering (2022) 4
60. Yang, Y., Soatto, S.: Fda: Fourier domain adaptation for semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4085–4095 (2020) 5
61. Zhai, X., Oliver, A., Kolesnikov, A., Beyer, L.: S4l: Self-supervised semi-supervised learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1476–1485 (2019) 1
62. Zhang, S., Khan, S., Shen, Z., Naseer, M., Chen, G., Khan, F.: Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 1, 4, 10