This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Generate, Refine, and Encode: Leveraging Synthesized Novel Samples for On-the-Fly Fine-Grained Category Discovery

Xiao Liu[1,3*], Nan Pu[2*†], Haiyang Zheng[2], Wenjing Li[1], Nicu Sebe[2], Zhun Zhong[1†]
[1]Hefei University of Technology, [2]University of Trento, [3]Helmholtz AI

Abstract

*In this paper, we investigate a practical yet challenging task: On-the-fly Category Discovery (OCD). This task focuses on the online identification of newly arriving stream data that may belong to both known and unknown categories, utilizing the category knowledge from only labeled data. Existing OCD methods are devoted to fully mining transferable knowledge from only labeled data. However, the transferability learned by these methods is limited because the knowledge contained in known categories is often insufficient, especially when few annotated data/categories are available in fine-grained recognition. To mitigate this limitation, we propose a diffusion-based OCD framework, dubbed Diff**GRE**, which integrates **G**eneration, **R**efinement, and **E**ncoding in a multi-stage fashion. Specifically, we first design an attribute-composition generation method based on cross-image interpolation in the diffusion latent space to synthesize novel samples. Then, we propose a diversity-driven refinement approach to select the synthesized images that differ from known categories for subsequent OCD model training. Finally, we leverage a semi-supervised leader encoding to inject additional category knowledge contained in synthesized data into the OCD models, which can benefit the discovery of both known and unknown categories during the on-the-fly inference process. Extensive experiments demonstrate the superiority of our DiffGRE over previous methods on six fine-grained datasets. The codes are available here: https://github.com/XLiu443/DiffGRE*

1. Introduction

Generalized category discovery (GCD), an emerging and challenging problem, has garnered increasing attention due to its potential for widespread real-world applications, such as discovering new diseases [57], bird species [29], and crops [23]. GCD aims to automatically cluster unlabeled data from unseen categories using knowledge from seen categories. However, current GCD methods [28, 29, 41, 47, 54] follow an offline inference paradigm, where category discovery is typically implemented by applying conventional clustering algorithms to a pre-collected batch of query data that requires discovery. This paradigm significantly limits the practicality of GCD techniques in real-world applications, where systems are expected to provide online feedback for each new instance.

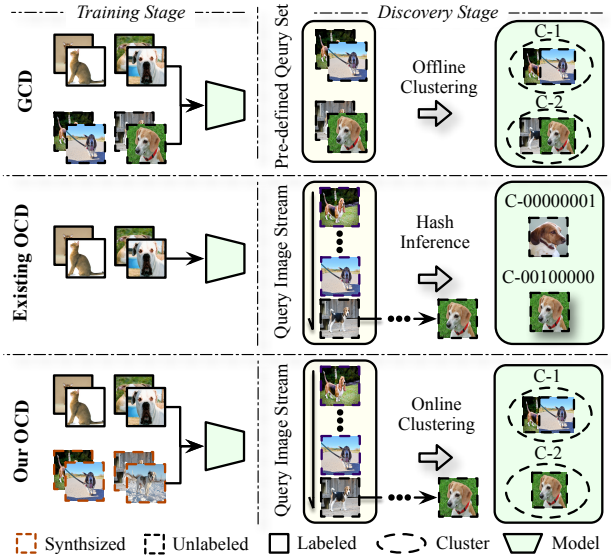


Figure 1. Comparison with GCD and OCD. Most existing GCD methods follow a transductive learning paradigm that requires previously accessing test data during training. Meanwhile, they utilize offline clustering to assign category labels for pre-collected query sets. In contrast, OCD models are trained with only labeled training data and can provide instance-wise inference for a stream of data that require discovery. Unlike existing OCD methods, our DiffGRE syntheses new training samples for OCD training, while performing more robust online cluster than hash inference.

To address this limitation, On-the-fly Category Discovery (OCD) [5] is introduced to eliminate the requirement of accessing pre-defined query sets in training phrases and enables GCD models to generate instance feedback for streaming data, as shown in Fig. 1. Existing OCD methods that focus on learning transferable knowledge from only labeled data and categories are often inefficient, as the knowl-

*Equal contribution.
†Corresponding author.

edge contained in known categories is often insufficient, especially in fine-grained scenarios where annotated data are often scarce. To address this limitation, we propose a diffusion-based OCD framework, namely DiffGRE, to synthesize new samples that contain additional category information and are helpful for fine-grained category discovery.

Diff**GRE** mainly includes **G**eneration, **R**efinement, and **E**ncoding stages. Specifically, most existing diffusion-based data augmentation methods [7, 13, 44] either heavily rely on fine-tuning the pre-trained diffusion model on large-scale target data or aim at faithfulness augmentation (*e.g.*, category-preserving generation [15, 38] or seen target-category argumentation [44]), which are not effective for the OCD task because 1) there has relatively few labeled data in OCD setting, especially for fine-grained datasets; 2) the data that require to discover include both known and unknown categories. However, we empirically find that augmenting only known-category data is struggling to benefit unknown-category discovery (see details in Appendix).

Considering that 1) diffusion models already have a semantic latent space [20, 32, 37] and 2) unseen categories can be synthesized by composing existing semantic attributes in seen categories [22, 24, 43], we propose to explore a cross-category latent-space interpolation that can synthesize virtual categories through semantic Attribute Composition Generation (ACG). ACG employs the latent interpolation strategy, which not only interpolates diffusion latent features but also integrates interpolation on the CLIP [30] embedding space. This enhances semantic information in the latent space, thereby enriching the attributes that can be combined. An example is shown in Fig. 2, where two images are used as inputs for ACG, resulting in a synthesized image. We interpolate between a "Laysan albatross" image and a "Black-footed albatross" image and synthesize a new bird image that belongs to a virtual category. Surprisingly, by searching from both known- and unknown-category data, we find that the nearest neighbor of the generated image belongs to an unknown category, *e.g.*, "Groove-billed Ani" in Fig. 2. "Groove-billed Ani" appears as a combination of the semantic attributes from "Laysan albatross" and "Laysan albatross", which shows the promising potential to synthesize unknown categories based on known-category attributes.

Although the image synthesis approach can achieve semantic attribute re-composition, it cannot ensure that each generated image is likely a new-category image and is helpful for OCD due to the randomness of diffusion process [14]. To mitigate this side-effect, we design a lightweight Diversity-Driven Refinement (DDR) to select the generated images that include diverse category information as much as possible. DDR filters off synthesized images that are visually similar to known categories, according to the average similarities between the image features and category-level representations. This mechanism not only improves the diversity of augmented samples but also reduces the computational consumption for subsequent model training.

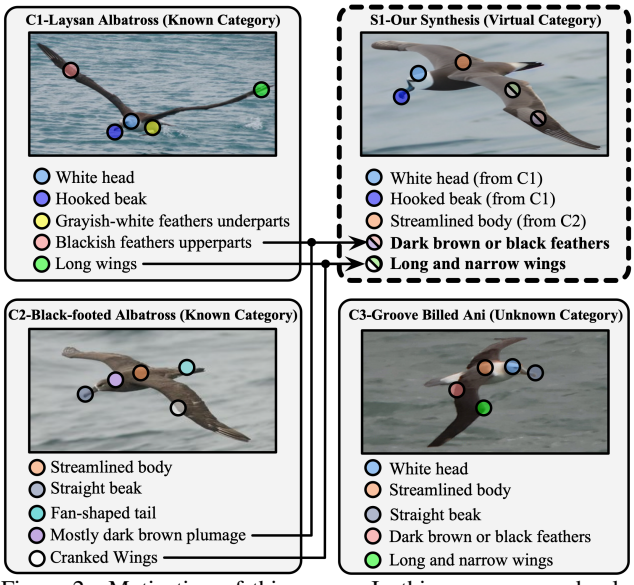


Figure 2. Motivation of this paper. In this case, we randomly sample two images from two known categories as inputs to synthesize a new image by attribute re-composition. We observe that the synthesized image has similar attributes to an unknown category defined in the OCD benchmark. This observation motivates us to fully leverage latent category attributes included in known categories to generate additional category information, thereby improving the existing OCD methods.

To take full advantage of the refined synthesized data, we propose a Semi-supervised Leader Encoding (SLE) to assign reliable pseudo labels for them. Importantly, the encoded leader representations in SLE are further used to implement accurate on-the-fly inference based on the high-dimensional features outputted from the backbone network. This differs from the existing method [5] that employs hash code as category descriptors, which inevitably degrades the representation ability of the high-dimensional features due to the binary nature. In contrast, our SLE can directly operate on original features containing rich discriminative information following an online clustering paradigm, which hence leads to better OCD performance.

Our contribution can be summarized as follows:

- We advocate embracing pre-trained and high-quality generative models based on the diffusion technique and introducing them into new category discovery tasks.
- We propose a new and plug-and-play DiffGRE framework to enhance the OCD abilities of existing methods.
- Extensive experiments verify the effectiveness of our DiffGRE on six fine-grained datasets.

2. Related Work

Novel Category Discovery (NCD), first introduced by DTC [11], aims to categorize unlabeled novel classes by transferring knowledge from labeled known classes. However, existing NCD methods [8, 11, 56] assume that all unlabeled data exclusively belong to novel classes. To release such an impractical assumption, Generalized Category Discovery (GCD) is proposed in [41], allowing unlabeled data to be sampled from both novel and known classes. While existing NCD/GCD methods [1, 3, 8, 17, 28, 41, 45, 45, 46, 52, 56] have shown promising results, two key assumptions still impede their real-world application. 1) These models heavily rely on a predetermined query set (the unlabeled dataset) during training, limiting their ability to handle truly novel samples and hindering generalization. 2) The offline batch processing of the query set during inference makes these models impractical for online scenarios where data emerges continuously and the model requires instant feedback. To address these limitations, Du et al. introduced the On-the-Fly Category Discovery (OCD) [5], which removes the assumption of a predefined query set and requires instance feedback with stream data input. They proposed the SMILE method, which identifies the category of each instance via a hash-descriptor inference. *Notably, compared with the above methods, our DiffGRE has two differences: 1) Unlike SMILE, which employs sensitive hash inference, DiffGRE directly leverages features for inference following an online clustering fashion. 2) Existing OCD methods utilize only labeled data for training. In contrast, our DiffGRE generates new samples and designs a new effective framework to exploit these synthesized data.*

Data Synthetic with Diffusion Models. Synthesizing new samples with diffusion models has been applied in various computer vision tasks, including image classification [15, 35, 39], domain adaptation [6, 16, 27] and zero/few-shot learning [4, 21, 48]. Most of these methods aim to augment images to enhance diversity while keeping faithfulness, which can be broadly divided into two groups. The one [38, 44] is to generate images based on Img2Img diffusion with pre-defined or learned text prompts. The other leverages Latent Perturbation [9, 58] to empower the ability of diffusion models to synthesize new samples. However, the former heavily depends on prompt quality and typically generates samples with high faithfulness and low diversity. In contrast, the latter produces more diverse images, but often with substantial noise. To mitigate these drawbacks, Diff-Mix [44] introduces a mix-up strategy that augments training data by transforming one known category into another. *However, these methods focus on generating new images of known categories, which is relatively ineffective for OCD that aims at new category discovery. To address this issue, we introduce DiffGRE to synthesize virtual category samples via attribute re-composition, tailor-made for OCD.*

3. Method

In this section, we first briefly introduce the OCD task and then present our approach, DiffGRE, which integrates Generation, Refinement, and Encoding in a multi-stage fashion.

Problem Definition. The setting of OCD is defined as follows. We are provided with a support set, denoted as $\mathcal{D}_S = \{(\mathbf{x}_i, y_i^s)\}_{i=1}^M \subseteq \mathcal{X} \times \mathcal{Y}_S$ for training, and a query set, denoted as $\mathcal{D}_Q = \{(\mathbf{x}_i, y_i^q)\}_{i=1}^N \subseteq \mathcal{X} \times \mathcal{Y}_Q$ for testing. Here, M and N represent the number of samples in $\mathcal{D}_S$ and $\mathcal{D}_Q$, respectively. $\mathcal{Y}_S$ and $\mathcal{Y}_Q$ are the label spaces for the support set and query set, respectively, where $\mathcal{Y}_S \subseteq \mathcal{Y}_Q$. We define classes in $\mathcal{Y}_S$ as known/old classes and classes in $\mathcal{Y}_Q/\mathcal{Y}_S$ as unknown/new classes. Only the support set $\mathcal{D}_S$ is used for model training. During testing, $\mathcal{D}_Q$ includes samples from both known and unknown categories, which are inferred one by one, allowing for online feedback.

Framework Overview. Here, we propose that DiffGRE synthesize new samples that include additional category knowledge to enhance the OCD abilities of existing methods. Fig. 3 shows an overview of DiffGRE's framework. In particular, DiffGRE consists of three components: attribute composition generation, diversity-driven refinements, and semi-supervised leader encoding.

3.1. Attribute Composition Generation

The Attribute Composition Generation (ACG) is proposed to synthesize virtual categories by interpolating the latent embeddings across different categories. The module is built based on previous Stable Diffusion Img2Img methods. We apply the latent interpolation strategy that performs interpolation in both the CLIP's [30] embedding space and the Stable Diffusion's latent space, facilitating attribute recomposition to synthesize novel images.

Stable Diffusion Img2Img. Different from vanilla diffusion models, the Stable Diffusion [32] efficiently performs image generation and manipulation in the latent space, synthesizing images by progressively refining the latent representations. Specifically, given an image I_i, the encoder $\mathcal{E}$ maps it to a latent z_i, *i.e.,* $z_i = \mathcal{E}(I_i)$. Besides, an additional input, *e.g.,* a text prompt t is fed to the model as a conditional input to control the synthesized image content. The text prompt is first pre-processed by a text tokenizer and then encoded by a text encoder, e.g., a pre-trained CLIP model.This can be formulated as: *i.e.,* $z_t = \mathcal{T}(t)$ where $\mathcal{T}$ denotes the text encoder of the Stable Diffusion model.

Latent Interpolation in ACG. Based on the above settings, Stable Diffusion Img2Img variants can be applied to edit images with text prompts. Previous works [10, 34, 44] in image editing have proposed using diffusion models to generate samples that have similar appearance attributes with known categories. However, this is incompatible with the OCD task. Our goal is to utilize the attributes of known categories to synthesize new samples belonging to virtual

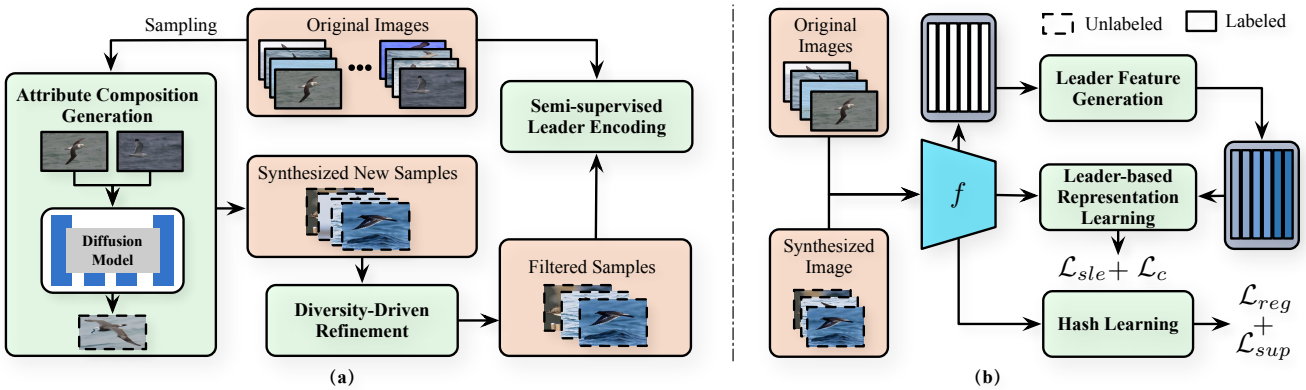


Figure 3. (a) Overview of our DiffGRE framework. First, we sample images from different categories to generate virtual-category images via our ACG. Then, we execute DDR to filter off the failure images and keep the remaining data for subsequential training. Last, we utilize SLE to exploit the generated data to further enhance the OCD models. (b) The pipeline of our SLE module. First, SLE mines the category-level relationship across the known and virtual categories for leader-based representation learning, following an iterative manner. After training, it leverages the generated leader feature as the known-category prior to performing robust online clustering inference.

categories. To achieve this, we build on the Stable Diffusion Img2Img model by incorporating semantic latent codes from two inputs for attribute re-composition.

In this work, we consider applying spherical interpolation [36] in three embedding spaces, *i.e.*, the latent space of Stable Diffusion, CLIP visual space, and CLIP textual space. Specifically, we use the VAE encoder in Stable Diffusion to extract the latent embeddings $\mathbf{z}^l$ and CLIP's visual encoder to extract visual embeddings $\mathbf{z}^v$. Meanwhile, we leverage a pre-trained image caption model [49] to transform images into textural descriptions. Then, we extract textural embedding $\mathbf{z}^t$ from the textural descriptions via CLIP textual encoder.

The overall interpolation processes are formulated as:

$$\overline{\mathbf{z}}^* = \frac{\sin((1-\lambda_*)\theta)}{\sin(\theta)}\mathbf{z}_1^* + \frac{\sin(\lambda_*\theta)}{\sin(\theta)}\mathbf{z}_2^*, \ * \in \{t, v, l\} \quad (1)$$

where $\theta = \arccos(\mathbf{z}_1^* \cdot \mathbf{z}_2^*)$ and $\lambda_* \in [0, 1]$ is the interpolation parameter, and $\mathbf{z}_1^*, \overline{\mathbf{z}}_2^*$ denote the latent embeddings of two inputs, respectively. Following Stable Diffusion [32], we leverage both textual and latent embeddings to generate noise that is subsequently used for the denoising process. Meanwhile, inspired by [25], we introduce CLIP visual embeddings to rectify the noise. The noise of the mixed latent is progressively removed at each timestep in the reverse denoising process, which is formulated as:

$$\begin{aligned} \mathcal{L}_{DM} &= \mathbb{E}_{t, \overline{\mathbf{z}}_\mathbf{0}^\mathbf{l}, \epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(\overline{\mathbf{z}}^\mathbf{l}, t, \overline{\mathbf{z}}^t\right)\right\|^2\right] \\ \overline{\mathbf{z}}_\mathbf{t}^\mathbf{l} &= \sqrt{\alpha_t}\overline{\mathbf{z}}_\mathbf{0}^\mathbf{l} + \sqrt{1-\alpha_t}\epsilon, \epsilon \sim \mathcal{N}(0, \mathbf{I}) \end{aligned} \quad (2)$$

where ϵ_θ is the UNet to predict noise, $\overline{\mathbf{z}}_\mathbf{0}^\mathbf{l}$ is the interpolated latent embedding of the input images, and α_t controls the noise schedule at timestep t. The final latent embedding is passed to the VAE decoder to obtain the synthesized image.

Discussion. In the OCD task, transforming an image from a known category into an unknown category based on a text prompt is challenging because textual information for new categories is not directly available. As a result, directly applying text-to-image editing or text-to-image generative models to synthesize new samples for the OCD task is not a feasible approach. Meanwhile, the latent space in Stable Diffusion primarily captures fine-grained visual details for image reconstruction, whereas the CLIP latent space is optimized for high-level semantic understanding. Integrating CLIP embeddings enables Stable Diffusion to interpolate diverse attributes, blending characteristics from multiple categories to synthesize virtual categories.

3.2. Diversity-Driven Refinement

Although ACG can generate numerous informative images, these images are not always favorable for OCD. As shown in Tab. 3, we experimentally find that direct training with all generated images disturbs discriminative feature extraction, especially for known categories. The primary reason may come from extensive and ambiguous synthesized samples that disrupt the original decision boundaries of known categories. To mitigate this issue, we propose the Diversity-Driven Refinement (DDR) module to retain samples that provide more additional information while filtering out those that do not, thereby reducing training overhead for the subsequent model training. Specifically, DDR removes samples that are too similar to known categories, preserving those that offer more diverse and valuable information.

First, we leverage the training samples to measure the quality of generated images. One straightforward way is calculating the cosine similarity between generated samples and all training images. However, this approach becomes unstable when image features follow a long-tailed distribution. Hence, we chose to compute all similarity scores between generated samples and category centers that can reliably capture the feature distribution of the original training samples in the embedding space. Moreover, we empirically

validate the effectiveness of the design in the Appendix.

Given the similarities, we set a threshold, denoted as γ, to filter out the images with relatively high scores. The DDR can be divided into the following steps:

1. Compute cosine similarity scores between generated images with original training category centers. The step is formulated as:

$$\mathrm{s}(z_{\text{gen}}, c_k) = \frac{z_{\text{gen}} \cdot c_k}{\|z_{\text{gen}}\| \|c_k\|}, \tag{3}$$

where z_{gen} represents the embedding of a generated image and c_k represents the center for category k.

2. For each generated sample, calculate the mean score over all class centers. It can be represented as:

$$s_{mean}(z_{\text{gen}}) = \frac{1}{K} \sum_{k=1}^{K} \mathrm{s}(z_{\text{gen}}, c_k), \tag{4}$$

where K is the total number of category centers.

3. Remove all images based on the threshold. Mathematically, it can be formulated as:

$$\mathcal{F}(I_{\text{gen}}) = \mathbf{1}(s_{mean}(z_{\text{gen}}) \leq \gamma) \cdot I_{\text{gen}}, \tag{5}$$

where $\mathcal{F}(I_{\text{gen}})$ is a function that decides whether to retain or remove the generated image I_{gen}.

3.3. Semi-Supervised Leader Encoding

Motivation. Although the combination of ACG and DDR is able to generate virtual category images, it is hard to assign precise category labels for each virtual image due to the randomness of attribute re-composition. We experimentally find that intuitively treating the generated images mixed by the same two categories as a unique new category leads to significant degradation of both known and unknown category discovery. Details are included in the Appendix.

Virtual Category Assignment. To fully mine the helpful knowledge from the generated virtual dataset, we propose a Semi-supervised Leader Encoding (SLE) approach to yield reliable category labels for virtual images. First, we combine the labeled data and generated data to build an agency training set $\mathcal{D}_A = \mathcal{D}_S \cup \mathcal{D}_G$. Secondly, we apply the off-the-shelf clustering method [33] on $\mathcal{D}_A$ and generate the virtual category labels $\mathcal{Y}_A$ for all data, resulting in $\mathcal{D}_A = \{(\mathbf{x}_i, y_i^s)\}_{i=1}^{M+A} \subseteq \mathcal{X} \times \mathcal{Y}_A$, where A is the number of synthesized image after DDR. Next, we leverage the ground-truth labels of partially-labeled data in $\mathcal{D}_A$ to rectify the clustering results further. Specifically, we first leverage the Hungarian optimal assignment algorithm [19] to find the optimal mapping between the initial clustering assignment and the ground-truth labels of partially-labeled data. Based on the mapping results, we replace the clustering assignments of both labeled and unlabeled generated data according to the ground-truth labels and yield a new aligned version of the clustering assignments. Eventually, we establish a rectified $\mathcal{D}_A$ for OCD training.

Leader Feature Generation. We calculate the average of the features that belong to the same category as a category-specific leader feature, $\mathbf{l}_m = \frac{1}{|\mathcal{Y}_A|} \sum_{\mathbf{x}_m \in \mathcal{Y}_A} f(\mathbf{x}_m)$, where the $|\mathcal{Y}_A|$ is the number of virtual categories. Accordingly, the leader set for $\mathcal{D}_A$ is defined by: $\mathcal{C}_A = \{\mathbf{l}_m\}_{m=1}^{|\mathcal{Y}_A|}$.

Leader-based Representation Learning. To encourage models to learn robust features with a strong ability to discriminate different categories, we propose leader-based contrastive learning to increase the inter-category distance while reducing the intra-category distance on the virtual categories data. The objective is formulated as:

$$\mathcal{L}_{sle}(\mathbf{x}_n, y_n) = -\log \frac{\exp\left(f(\mathbf{x}_n) \cdot \mathbf{l}_{y_n}^{\mathrm{T}} / \tau\right)}{\sum_{m=1, m \neq y_n}^{|\mathcal{Y}_\mathcal{A}|} \exp\left(f(\mathbf{x}_n) \cdot \mathbf{l}_m^{\mathrm{T}} / \tau\right)}, \tag{6}$$

where τ is the temperature factor and is empirically set to 0.05 for all experiments.

Furthermore, inspired by the success of a combination of contrastive loss and classification loss in the other open-set task [53], we employ an additional classifier to supplement the leader-based contrastive learning with a cross-entropy loss, denoted as $\mathcal{L}_c$.

Remark. Our SLE is tailor-made for the OCD task and brings three advantages: 1) SLE provides an effective manner to leverage the generated image with less noise, which injects additional category knowledge into OCD models, thereby resulting in better performance. 2) Unlike the existing OCD methods that apply representation learning on a low-dimensional hash space, our SLE reduces the information loss by performing contrastive learning on a high-dimensional feature space. 3) The calculated leader features can be stored as known-category prior, benefiting the subsequent on-the-fly inference in Sec. 3.4.

3.4. Optimization and Inference

Model Training. We add our DiffGRE to the existing OCD method SMILE [5]. The SMILE method applies two hash heads to implement the OCD task. In order to enhance the representation embedding of old classes, we add another classifier to the final loss function. The total loss is:

$$\mathcal{L} = \mathcal{L}_{sup} + \mathcal{L}_{reg} + \alpha * \mathcal{L}_{sle} + \beta * \mathcal{L}_c, \tag{7}$$

where L_{sup} and L_{reg} are basic losses from SMILE [5]. Here, L_{sup} is utilized for supervised contrastive learning, while L_{reg} serves to constrain the output of the hash head $\mathcal{H}_h(\cdot)$ in [5] to approximate discrete values of +1 or -1. And α and β weight the two losses to balance their contributions.

Discussion. Existing OCD methods use hash-like mechanisms, which suffer from sensitivity, particularly for new categories. This is demonstrated by the results in Tab. 1, where the hash-based methods are shown to be more sensitive compared to PHE [55]. To address this, we design Online Cluster Inference (OCI) with high-dimensional features, which preserves rich information while fully leveraging the learned leader representation in SLE and mitigating the sensitivity issues.

Online Cluster Inference based on SLE. In the testing stage, we first establish a dynamic leader memory, which is initialized by the category-specific leader features in Sec. 3.2. Then, we calculate the maximal intra-category distance, Δ_{max}, as an adaptive threshold to determine if this instance belongs to unknown categories following an online clustering fashion. Given a test instance, $\mathbf{x}_j \in \mathcal{D}_Q$, if it is predicted to belong to known categories, we assign its estimated category label $\hat{y}_j$ to the category label according to the nearest leader. Otherwise, we create a new leader with the instance feature and append it to the dynamic leader memory. During the testing process, we update both known- and unknown-category leaders by momentum averaging of corresponding instance features. The algorithm is described in the Appendix.

4. Experiments

4.1. Experiment Setup

Datasets. We conduct experiments on six fine-grained datasets, including CUB-200 [42], Stanford Cars [18], Oxford-IIIT Pet [26], and three super-categories from the more challenging dataset, iNaturalist [40], including Arachnida, Animalia, and Mollusca. Details about the datasets are provided in the Appendix.

Evaluation Metrics. We follow [5] and adopt clustering accuracy as the evaluation metric, $ACC = \frac{1}{|\mathcal{D}_Q|}\sum_{i=1}^{|\mathcal{D}_Q|}\mathbb{I}(y_i = C(\hat{y}_i))$, where $\hat{y}_i$ represents the predicted labels and y_i denotes the ground truth. The function C denotes the optimal permutation that aligns predicted cluster assignments with the actual class labels. We report "ACC-ALL", "ACC-OLD" and "ACC-NEW" for all categories, known categories, and unknown categories.

Implementation Details. For a fair comparison, we follow OCD [5] and use the DINO-pretrained ViT-B-16 [2] as the backbone. During training, only the final block of ViT-B-16 is fine-tuned. The Projector $\mathcal{H}_h(\cdot)$ consists of three linear layers with an output dimension set to $L = 12$, which produces $2^{12} = 4096$ binary category encodings. We follow OCD to set this dimension for a fair comparison. The ratio α and β in the total loss are set to 0.3 and 1.0 for all datasets. More details about γ analysis are provided in Fig. 4.

Inference Strategy. Existing OCD methods mainly employ the *hash-like inference* strategy, where image features are mapped to hash-like category descriptors, and images sharing the same descriptor are grouped into a class to enable on-the-fly inference. However, this approach suffers from accuracy degradation due to binarization and limited hash length. In this work, we introduce the Online Clustering Inference (OCI) strategy based on our SLE. The OCI continuously measures similarities between streaming inputs and category-specific leader features, dynamically assigning new inputs to known categories or creating new categories for them.

4.2. Comparison with State-of-the-Art Methods

As the proposed SLE module enables hash-like inference methods to tackle OCD via Online Clustering Inference (OCI), we denote testing SMILE with SLE-based OCI strategy as "SMILE + SLE-based". Meanwhile, "SMILE + DiffGRE" can achieve on-the-fly inference in both hash-like and OCI strategies. For a thorough and fair comparison, we conduct detailed experiments under the two inference strategies to verify our plug-and-play DiffGRE.

Results of Hash-like Inference. The results compared with hash-like inference methods are shown in Tab. 1, where it is obvious that the introduction of the DiffGRE framework significantly enhances the performance of these strong baselines across all metrics. For example, incorporating DiffGRE yields average improvements of 6.5% in ACC-ALL and 11.5% in ACC-OLD for BaseHash across six datasets. Similarly, SMILE demonstrates average gains of 3.7% in ACC-ALL and 7.6% in ACC-OLD. These promising results demonstrate that DiffGRE effectively enhances on-the-fly category discovery for existing hash-like inference methods. Moreover, "PHE + DiffGRE" consistently outperforms all competitors, achieving the highest ACC-ALL averages.

Results of Online Clustering Inference. The results compared with Online Clustering Inference methods are shown in Tab. 2. Our SLE enables hash-like inference methods to perform OCI and achieves higher accuracy than the leading competitor SLC. For example, "BaseHash + SLE-based" exceeds "BaseHash + SLC" by an average of 10.8% in ACC-ALL and 16.5% in ACC-New across six datasets. Moreover, the incorporation of our full DiffGRE framework consistently enhances all methods. Specifically, it boosts "BaseHash + SLE-based" by an average of 2.1% in ACC-ALL, and enhances "SMILE + SLE-based" by an average of 1.6% in ACC-ALL. Although DiffGRE improves PHE in "PHE + DiffGRE" setting, the improvement is relatively lower than that under the hash-like inference strategy. This is because PHE focuses on optimizing a superior hash network, which limits the training of the feature extractor. Consequently, the inference improvements in OCI strategy are limited by suboptimal features.

Comprehensive evaluation. Our DiffGRE consistently improves performance across all datasets under both hash-like and OCI inference strategies. Notably, it achieves state-of-the-art average accuracy over six datasets with the OCI strategy, supporting our discussion in Sec. 3.4.

4.3. Ablation Study

We conduct an ablation study of DiffGRE on the Arachnida, Mollusca, and CUB datasets using hash-like inference, as shown in Tab. 3. "SMILE + DiffGRE" refers to the full method with DiffGRE applied to SMILE. Removing the SLE module ("$w/o\ \mathcal{L}_{sle}$") leads to an average reduction of 3.6% in ACC-ALL across the three datasets. This confirms that the SLE module effectively introduces additional cat-

Table 1. Comparison with **hash-like inference** methods, with the best results in **bold** and the second best underlined.

Method	Arachnida			Mollusca			Oxford Pets			Average		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New
BaseHash [5]	23.6	42.3	10.8	25.7	40.0	18.0	35.8	36.1	35.7	28.4	39.5	21.5
BaseHash [5] + DiffGRE (Ours)	27.2	50.5	12.4	28.8	40.6	20.9	38.6	39.2	39.2	31.5	43.4	24.2
SMILE [5]	27.9	53.0	12.2	33.5	42.8	28.6	41.2	42.1	40.7	34.2	46.0	27.2
SMILE [5] + DiffGRE (Ours)	35.4	66.8	**15.6**	36.5	44.2	**32.5**	42.4	42.1	42.5	38.2	51.0	**30.2**
PHE [55]	37.0	75.7	12.6	39.9	**65.0**	26.5	48.3	**53.8**	45.4	41.7	**64.8**	28.2
PHE [55] + DiffGRE (Ours)	**39.1**	**76.5**	**15.6**	**40.3**	63.4	28.1	**48.6**	52.6	**46.6**	**42.7**	64.2	30.1
Method	CUB			Stanford Cars			Animalia			Average		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New
BaseHash [5]	21.1	26.2	18.5	15.4	23.0	11.7	31.2	49.7	23.6	22.6	33.0	17.9
BaseHash [5] + DiffGRE (Ours)	34.2	48.8	26.9	26.9	47.3	**17.0**	36.4	60.2	26.6	32.5	52.1	23.5
SMILE [5]	32.2	50.9	22.9	26.2	46.7	16.3	34.8	58.8	24.9	31.1	52.1	21.4
SMILE [5] + DiffGRE (Ours)	35.4	**58.2**	23.8	30.5	59.3	16.5	37.4	**69.3**	24.3	34.4	**62.3**	21.5
PHE [55]	36.4	55.8	27.0	31.3	61.9	16.8	40.3	55.7	31.8	36.0	57.8	25.2
PHE [55]+ DiffGRE (Ours)	**37.9**	57.0	**28.3**	**32.1**	**63.3**	16.9	**42.0**	64.5	**32.6**	**37.3**	61.6	**25.9**

Table 2. Comparison with **online clustering inference** methods, with the best results in **bold** and the second best underlined.

Method	Arachnida			Mollusca			Oxford Pets			Average		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New
BaseHash [5] + SLC [12]	25.4	44.6	11.4	31.1	59.8	15.0	35.5	41.3	33.1	30.7	48.6	19.8
BaseHash [5] + SLE-based (Ours)	42.0	55.0	33.9	40.7	49.6	36.0	45.5	48.0	44.0	42.7	50.9	38.0
BaseHash [5] + DiffGRE (Ours)	44.9	60.7	34.9	41.0	49.0	36.7	46.3	43.0	48.1	44.1	50.9	39.9
SMILE [5] + SLE-based (Ours)	43.7	60.0	37.0	42.7	54.3	36.6	49.1	49.1	49.1	45.2	54.5	40.9
SMILE [5] + DiffGRE (Ours)	45.3	56.6	**40.6**	**43.6**	51.0	**39.6**	**51.1**	**54.7**	49.2	**46.7**	54.1	**43.1**
PHE [55] + SLE-based	44.9	**78.9**	23.5	40.3	60.8	29.5	46.2	44.4	47.1	43.8	61.4	33.4
PHE [55] + DiffGRE (Ours)	**47.7**	76.6	29.4	42.6	**62.0**	32.3	49.6	50.1	**49.3**	46.6	**62.9**	37.0
Method	CUB			Stanford Cars			Animalia			Average		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New
BaseHash [5] + SLC [12]	28.6	44.0	20.9	14.0	23.0	9.7	32.4	61.9	19.3	25.0	43.0	16.6
BaseHash [5] + SLE-based (Ours)	37.3	38.4	36.7	21.7	30.3	17.5	44.9	57.3	39.8	34.6	42.0	31.3
BaseHash [5] + DiffGRE (Ours)	38.5	40.6	37.5	26.0	37.9	**20.3**	47.4	59.6	42.3	37.3	46.0	33.4
SMILE [5] + SLE-based (Ours)	39.3	43.4	37.2	27.5	47.4	18.0	48.8	61.0	43.8	38.5	50.6	33.0
SMILE [5] + DiffGRE (Ours)	**42.8**	47.8	**40.2**	**28.3**	**49.3**	18.1	**49.2**	61.1	**44.3**	**40.1**	**52.7**	**34.2**
PHE [55]+ SLE-based (Ours)	41.0	52.1	35.5	25.2	44.5	15.9	40.5	61.6	32.9	35.6	**52.7**	28.1
PHE [55] + DiffGRE (Ours)	42.5	**54.4**	36.5	27.7	48.1	17.8	43.5	**63.2**	35.3	37.9	55.2	29.9

Table 3. Ablation study on training components.

Method	Arachnida			Mollusca			CUB		
	All	Old	New	All	Old	New	All	Old	New
SMILE [5]	27.9	53.0	12.2	33.5	42.8	28.6	32.2	50.9	22.9
$w/o\ \mathcal{L}_{sle}$	29.3	53.7	13.9	33.9	41.4	29.9	33.2	53.9	22.9
$w/o\ \mathcal{L}_{c}$	34.5	66.3	14.6	36.0	42.9	32.3	33.6	55.4	22.7
w/o DDR	33.5	62.8	15.2	34.5	39.4	31.9	32.4	53,2	22.0
SMILE + DiffGRE	**35.4**	**66.8**	**15.6**	**36.5**	**44.2**	**32.5**	**35.4**	**58.2**	**23.8**

egory knowledge into the training process. Removing the DDR module leads to an average drop of 2.3% in ACC-ALL and 4.6% in ACC-OLD across the datasets, indicating that low-quality samples synthesized by diffusion models can disrupt discriminative feature extraction. This also underscores the effectiveness of the DDR module in filtering out low-quality samples. In addition, compared to SMILE, the "w/o DDR" variant shows an average improvement of 3.2% across the datasets. These findings suggest that supplementing with additional category information, even if noisy, can improve the identification of fine-grained categories and enhance the overall effectiveness of category discovery.

4.4. Analysis

Hyper-Parameters Analysis. 1) The impact of the ratios α and β in Eq. (7) is illustrated in Fig. 4 (a-b). Parameters α and β control the relative importance of L_{sle} and L_c during the training process, respectively. The parameter α is relatively insensitive, showing subtle variation in accuracy when adjusted from 0.1 to 0.9. We fine-tune this parameter on the CUB dataset and fixed it at 0.3 for all datasets. Similarly, β remains stable across variations, as shown in Fig. 4 (b). We set $\beta = 1.0$ for all datasets during training. 2) We examine the impact of γ in the DDR module on the Arachnida dataset. This parameter controls the filtering of synthesized samples, where a lower γ results in fewer samples being filtered. As shown in Fig. 4 (c) and (d), when the number of remaining images is comparable to that of labeled data, the model reaches optimal accuracy on unknown categories. Thus, we set 0.4 for the Arachnida dataset. Similarly, we set γ for other datasets according to their number of labeled data. Moreover, we explore the impact of interpolation parameters λ_* in Eq. (1), where $* \in \{t, v, l\}$. We

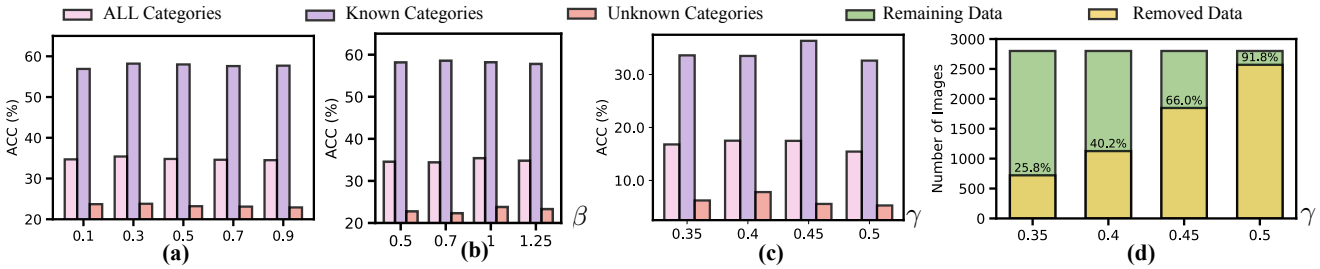


Figure 4. Impact of hyperparameters, *i.e.,* α, β and γ. We implement experiments on CUB [42] in **(a-b)** and Arachnida [40] in **(c-d)**.

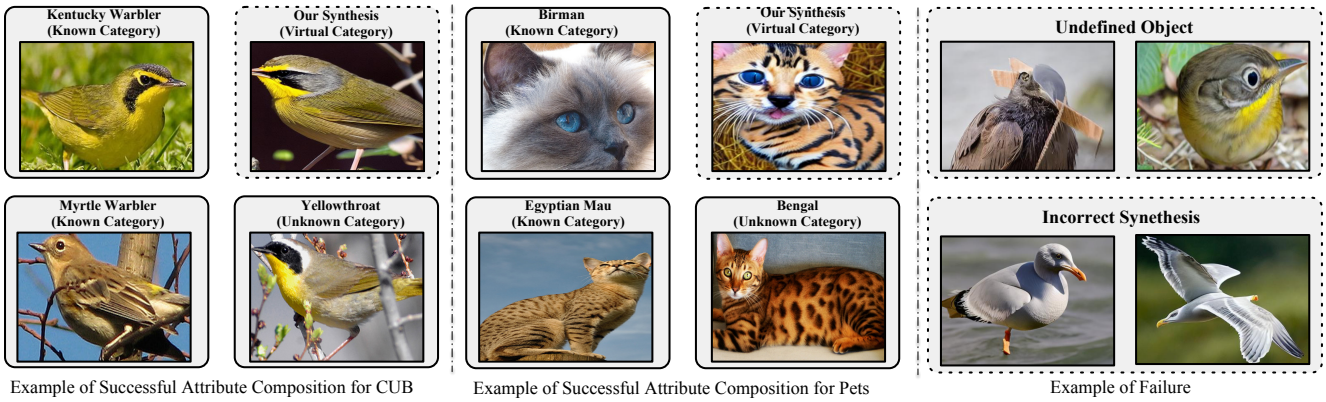


Figure 5. Visualization of the synthesized images. We illustrate two examples of successful attribute composition on the CUB [42] and Pets [26] datasets. Each successful example includes a synthesized sample and a real sample belonging to an unknown category, which is extracted through the nearest neighbor search. The nearest neighbor search is performed on both the known and unknown category images with respect to the synthesized samples. We show examples of failed generations: *undefined object* and *incorrect synthesis*.

follow the previous work [31] to set $\lambda_t = 0.7$. Then, we fix λ_t to tune the optimal $\lambda_v = 0.7$ and $\lambda_l = 0.8$ (see detailed analysis in the Appendix).

Impact of Different Synthesis Methods. To further validate our method, we compare it with two alternative sample synthesis strategies. The first includes T2I generative models, Diff-Mix [44] and Da-Fusion [38], using known categories as text inputs. For Diff-Mix, we also use unknown categories as text inputs, though this violates the OCD task constraints. The second strategy adopts pixel-level mixing methods, CutMix [50] and MixUp [51]. Both are used as replacements for our ACG to ensure a fair comparison. The results are summarized in Tab. 4. As shown, pixel-level methods underperform compared to ours. While T2I models generate high-quality images, their improvement for unknown categories is limited, proving that Diff-Mix can only synthesize images within known categories, making it inefficient for the OCD task. In contrast, our method achieves superior accuracy, validating its effectiveness in synthesizing novel samples for unknown categories.

4.5. Visualization and Limitation

Visualization. We visualize the generated images on different datasets in Fig. 5. Qualitatively, images generated by our ACG show clear attribute features while preserving the primary characteristic of the objects. Notably, the nearest neighbor of the generated image often belongs to the unknown categories in the OCD benchmark. We also present several failure cases, categorizing outputs with noticeable errors or defects as "incorrect", and those that are entirely unrecognizable as "undefined". These failures are mainly caused by the randomness of the diffusion models (see more examples and comparisons in the Appendix).

Table 4. Impact of different synthesis methods.

Method	CUB		
	All	Old	New
SMILE [5]	32.2	50.9	22.9
SMILE [5] + Da-Fusion [38]	34.4	56.7	23.2
SMILE [5] + Diff-Mix [44] (w/ known category)	32.9	54.4	22.1
SMILE [5] + Diff-Mix [44] (w/ unknown category)	35.1	56.6	24.3
SMILE [5] + CutMix [50]	32.8	54.6	21.9
SMILE [5] + MixUp [51]	33.3	55.6	22.1
SMILE [5] + DiffGRE (Ours)	35.4	58.2	23.8

5. Conclusion

This work addresses the practical and challenging OCD task. To enhance existing OCD models with limited category knowledge, we propose to embrace pre-trained diffusion-based generative models and design a plug-and-play DiffGRE framework to synthesize new training samples. DiffGRE aims to generate new samples with additional category knowledge and effectively leverage these images to enhance OCD models. Extensive experiments on six datasets validate the effectiveness of our approach.

Acknowledgement

This work was funded by the National Natural Science Foundation of China (No. 62402157) and the Fundamental Research Funds for the Central Universities (No. JZ2025HGTB0219). This work was partly supported by the MUR PNRR project FAIR (PE00000013) and funded by the NextGenerationEU, the EU Horizon projects ELIAS (No. 101120237) and ELLIOT (No. 101214398). The computation is completed on the HPC Platform of Hefei University of Technology.

References

[1] Wenbin An, Feng Tian, Qinghua Zheng, Wei Ding, QianYing Wang, and Ping Chen. Generalized category discovery with decoupled prototypical network. In *AAAI*, 2023. 3

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 6

[3] Florent Chiaroni, Jose Dolz, Ziko Imtiaz Masud, Amar Mitiche, and Ismail Ben Ayed. Parametric information maximization for generalized category discovery. In *ICCV*, 2023. 3

[4] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. *NeurIPS*, 2024. 3

[5] Ruoyi Du, Dongliang Chang, Kongming Liang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. On-the-fly category discovery. In *CVPR*, 2023. 1, 2, 3, 5, 6, 7, 8

[6] Zhekai Du and Jingjing Li. Diffusion-based probabilistic uncertainty estimation for active domain adaptation. *NeurIPS*, 2023. 3

[7] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *ICCV*, 2023. 2

[8] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *ICCV*, 2021. 3

[9] Yunxiang Fu, Chaoqi Chen, Yu Qiao, and Yizhou Yu. Dreamda: Generative data augmentation with diffusion models. *arXiv preprint arXiv:2403.12803*, 2024. 3

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[11] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*, 2019. 3

[12] John A Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975. 7

[13] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022. 2

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2

[15] Khawar Islam, Muhammad Zaigham Zaheer, Arif Mahmood, and Karthik Nandakumar. Diffusemix: Label-preserving data augmentation with diffusion models. In *CVPR*, 2024. 2, 3

[16] Gwanghyun Kim, Ji Ha Jang, and Se Young Chun. Podia-3d: Domain adaptation of 3d generative model across large domain gap using pose-preserved text-to-image diffusion. In *ICCV*, 2023. 3

[17] Hyungmin Kim, Sungho Suh, Daehwan Kim, Daun Jeong, Hansang Cho, and Junmo Kim. Proxy anchor-based unsupervised learning for continuous generalized category discovery. In *ICCV*, 2023. 3

[18] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshop*, 2013. 6

[19] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 1955. 5

[20] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have A semantic latent space. In *ICLR*, 2023. 2

[21] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *ICCV*, 2023. 3

[22] Yun Li, Zhe Liu, Hang Chen, and Lina Yao. Context-based and diversity-driven specificity in compositional zero-shot learning. In *CVPR*, 2024. 2

[23] Yu Liu, Yaqi Cai, Qi Jia, Binglin Qiu, Weimin Wang, and Nan Pu. Novel class discovery for ultra-fine-grained visual categorization. In *CVPR*, 2024. 1

[24] Yu Liu, Jianghao Li, Yanyi Zhang, Qi Jia, Weimin Wang, Nan Pu, and Nicu Sebe. Pmgnet: Disentanglement and entanglement benefit mutually for compositional zero-shot learning. *CVIU*, 249:104197, 2024. 2

[25] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 4

[26] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 6, 8

[27] Duo Peng, Qiuhong Ke, ArulMurugan Ambikapathi, Yasin Yazici, Yinjie Lei, and Jun Liu. Unsupervised domain adaptation via domain-adaptive diffusion. *IEEE TIP*, 2024. 3

[28] Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *CVPR*, 2023. 1, 3

[29] Nan Pu, Wenjing Li, Xingyuan Ji, Yalan Qin, Nicu Sebe, and Zhun Zhong. Federated generalized category discovery. In *CVPR*, 2024. 1

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3

[31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 8

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 4

[33] Martin Rosvall, Daniel Axelsson, and Carl T Bergstrom. The map equation. *The European Physical Journal Special Topics*, 2009. 5

[34] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 3

[35] Jordan Shipard, Arnold Wiliem, Kien Nguyen Thanh, Wei Xiang, and Clinton Fookes. Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion. In *CVPR*, 2023. 3

[36] Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pages 245–254, 1985. 4

[37] Zitao Shuai, Chenwei Wu, Zhengxu Tang, Bowen Song, and Liyue Shen. Latent space disentanglement in diffusion transformers enables zero-shot fine-grained semantic editing. *arXiv preprint arXiv:2408.13335*, 2024. 2

[38] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023. 2, 3, 8

[39] Cheng-Chang Tsai, Yuan-Chih Chen, and Chun-Shien Lu. Test-time stain adaptation with diffusion models for histopathology image classification. In *ECCV*, 2025. 3

[40] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018. 6, 8

[41] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *CVPR*, 2022. 1, 3

[42] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *Computation & Neural Systems Technical Report*, 2011. 6, 8

[43] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *NeurIPS*, 2023. 2

[44] Zhicai Wang, Longhui Wei, Tan Wang, Heyu Chen, Yanbin Hao, Xiang Wang, Xiangnan He, and Qi Tian. Enhance image classification via inter-class image mixup with diffusion model. In *CVPR*, 2024. 2, 3, 8

[45] Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *ICCV*, 2023. 3

[46] Yanan Wu, Zhixiang Chi, Yang Wang, and Songhe Feng. Metagcd: Learning to continually learn in generalized category discovery. In *ICCV*, 2023. 3

[47] Fengxiang Yang, Nan Pu, Li Wenjing, Luo Zhiming, Li Shaozi, Niculae Sebe, and Zhun Zhong. Learning to distinguish samples for generalized category discovery. In *ECCV*, 2024. 1

[48] Zebin You, Yong Zhong, Fan Bao, Jiacheng Sun, Chongxuan Li, and Jun Zhu. Diffusion models and semi-supervised learners benefit mutually with few labels. *NeurIPS*, 2024. 3

[49] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 4

[50] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 8

[51] Hongyi Zhang, YND Moustapha Cisse, and David Lopez-Paz. mixup: Beyond 903 empirical risk minimization. In *ICLR*, 2018. 8

[52] Bingchen Zhao and Oisin Mac Aodha. Incremental generalized category discovery. In *ICCV*, 2023. 3

[53] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In *CVPR*, 2019. 5

[54] Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Textual knowledge matters: Cross-modality co-teaching for generalized visual class discovery. In *ECCV*, 2024. 1

[55] Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Prototypical hash encoding for on-the-fly fine-grained category discovery. *NeurIPS*, 2025. 5, 7

[56] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *CVPR*, 2021. 3

[57] Jiaying Zhou, Yang Liu, and Qingchao Chen. Novel class discovery in chest x-rays via paired images and text. In *AAAI*, 2024. 1

[58] Yongchao Zhou, Hshmat Sahak, and Jimmy Ba. Training on thin air: Improve image classification with generated data. *arXiv preprint arXiv:2305.15316*, 2023. 3