Hyperbolic Category Discovery
–*Supplementary Material*–

Yuanpei Liu* Zhenqi He* Kai Han†
Visual AI Lab, The University of Hong Kong
{ypliu0,zhenqi_he}@connect.hku.hk kaihanx@hku.hk

We provide additional details and experimental results in this supplementary material, which is organized as follows:

- §A More Experimental Details
- §B More Quantitative Results
- §C More Qualitative Results

A. More Experimental Details

A.1. Dataset Statistics

For each dataset, we adhere to the data splitting scheme described in [16]. In this scheme, 50% of the classes will be sampled as 'Old', with the exception of CIFAR-100, which samples 80% of the classes. Following this, 50% of the images from known classes are used to create the labelled dataset $\mathbf{D}_l$, while the remaining images are allocated to the unlabelled dataset $\mathbf{D}_u$. The statistics for all the datasets utilized in this work are summarized in Tab. 1.

Table 1. Overview of the dataset, including the classes in the labelled and unlabelled sets ($M = |\mathbf{Y}_l|$, $K = |\mathbf{Y}_l \cup \mathbf{Y}_u|$) and counts of images ($|\mathbf{D}_l|$, $|\mathbf{D}_u|$). 'FG' denotes fine-grained.

Dataset	FG	$\lvert\mathbf{D}_l\rvert$	M	$\lvert\mathbf{D}_u\rvert$	K
CIFAR-10 [7]	✗	12.5K	5	37.5K	10
CIFAR-100 [7]	✗	20.0K	80	30.0K	100
ImageNet-100 [3]	✗	31.9K	50	95.3K	100
CUB [19]	✓	1.5K	100	4.5K	200
Stanford-Cars [6]	✓	2.0K	98	6.1K	196
FGVC-Aircraft [8]	✓	1.7K	50	5.0K	100
Herbarium19 [15]	✓	8.9K	341	25.4K	683
Oxford-Pet [10]	✓	0.9K	19	2.7K	37

A.2. Additional Implementation Details

Consistent with prior studies [13, 16, 21], we employ the ViT-B architecture [4] with pretrained weights from either DINO [2] or DINOv2 [9] as our backbone network. For our proposed hyperbolic methods, we adhere to nearly all hyperparameter settings established in [13, 16, 21] to facilitate fair comparisons with their respective baselines. The specific details are summarized as follows: For Hyp-SimGCD and Hyp-GCD, only the last block of the backbone is fine-tuned across all datasets. In contrast, Hyp-SelEx implements dataset-specific fine-tuning: the last two blocks are fine-tuned for CUB [19], FGVC-Aircraft [8], and all generic datasets, while the last three blocks are fine-tuned for Stanford-Cars [6]. Regarding method-specific hyperparameters, for Hyp-SimGCD, we set the weight ξ, which controls the weight of mean entropy loss, to 1.0 for all the datasets. For Hyp-SelEx, we follow [13] in setting α, which regulates label smoothing, to 0.5 for FGVC-Aircraft [8], 1.0 for CUB [19] and Stanford-Cars [6], and 0.1 for generic datasets. Additionally, the proposed parameter α_d, which balances distance-based and angle-based losses, linearly increases from 0 to its maximum value during training

*Equal contribution.
†Corresponding author.

according to the formula: $\alpha_d = \frac{e*\alpha_d^{\max}}{200}$, where e is the current training epoch. Specifically, we set $\alpha_d^{\max}$ to 1 for fine-grained and 0.5 for generic datasets.

A.3. Details of Hyp-SelEx

[13] proposes a hierarchical non-parametric method, SelEx, to address fine-grained GCD through a novel concept of *self-expertise*. It begins by constructing hierarchical pseudo-labeling via a *balanced semi-supervised k-means* algorithm to initialize clusters for known categories and then iteratively refines them by incorporating an equal number of random samples for unseen categories to balance cluster distribution. Following it, *supervised self-expertise* leverages weakly-supervised pseudo labels to group samples by capturing abstract-level similarity, whereas *unsupervised self-expertise* focuses on distinguishing semantically similar hard negative samples within the same clusters to sharpen fine-grained categorization.

Its representation learning objective composes of unsupervised self-expertise loss $\mathcal{L}_{\text{USE}}$ and supervised self-expertise loss $\mathcal{L}_{\text{SSE}}$. The unsupervised self-expertise loss, defined as $\mathcal{L}_{\text{USE}} = \ell_{ce}(\mathbf{p}, \hat{\mathbf{t}})$, calculates the binary cross entropy loss between the logits $\mathbf{p}$ and an adjusted target $\hat{\mathbf{t}}$, where $\mathbf{p}$ is calculated based on Euclidean distance, unlike prior GCD [16] approach that utilizes cosine similarity. [13] introduces an adjusted target matrix $\hat{\mathbf{t}}$ to recalibrate targets based on semantic similarity between samples. Specifically, $\hat{\mathbf{t}} = \alpha\mathbf{t} + (1-\alpha)\mathbf{I}$, where $\mathbf{t}$ can be calculated using $\mathbf{t} = [\sum_{k=1}^{\lg K} \frac{\mathbb{1}(\hat{y}_i^k = \hat{y}_j^k)}{2^k}]$ based on pseudo label $\hat{y}_i^k$ and $\hat{y}_j^k$ from hierarchical level k. α is the hyperparameter to control the label smoothing by identity metric $\mathbf{I}$. Then, the hierarchical supervised self-expertise loss can be denoted as:

$$\mathcal{L}_{\text{SSE}} = \frac{1}{2}\left(\sum_{k=0}^{\lg K} \frac{\mathcal{L}_s^k|\frac{\mathbf{d}}{2^k}}{2^k}\right), \tag{1}$$

where $\mathcal{L}_s^k|\frac{\mathbf{d}}{2^k}$ represents the supervised representation loss applied exclusively to the segment $\frac{\mathbf{d}}{2^k}$ of the embedding vector $\mathbf{d}$, corresponding to each level of the hierarchy. The final representation loss is given by $\mathcal{L}_{rep} = (1-\lambda_b)\mathcal{L}_{\text{USE}} + \lambda_b\mathcal{L}_{\text{SSE}}$. To combine SelEx with hyperbolic embeddings, we extend the hierarchical representation learning used in [13] into the hyperbolic space, utilizing the methodology introduced in the main paper.

Following the above pace, our Hyp-SelEx utilizes hyperbolic supervised and unsupervised self-expertise, denoted as $\mathcal{L}_{\text{SSE}}^{\mathbb{H}}$ and $\mathcal{L}_{\text{USE}}^{\mathbb{H}}$, respectively. Given two randomly augmented views $\mathbf{x}_i$ and $\mathbf{x}_i'$ for the same image in a mini-batch B, $\mathbf{z}_i$ and $\mathbf{z}_i'$ represent the feature extracted from backbone network ϕ and projector ρ_r of these two views in the Euclidean space, represented as $\mathbf{z}_i = \rho_r(\phi(\mathbf{x}_i))$. As introduced in Sec.3.4 of the main paper, we employ a hybrid of distance-based and angle-based loss functions, and hence the unsupervised self-expertise loss is represented as:

$$\mathcal{L}_{\text{USE}}^{\mathbb{H}} = \alpha_d \ell_{ce}(\mathbf{p}_{\text{dis}}, \hat{\mathbf{t}}) + (1-\alpha_d)\ell_{ce}(\mathbf{p}_{\text{ang}}, \hat{\mathbf{t}}), \tag{2}$$

where $\mathbf{p}_{\text{dis}}$ is the logit calculated based on the negative hyperbolic distance, expressed as $\mathcal{S}_d(\mathcal{M}(\mathbf{z}_i), \mathcal{M}(\mathbf{z}_i'))$, and $\mathbf{p}_{\text{ang}}$ is the logit calculated based on the original distance metrics, expressed as $\mathcal{S}_a(\mathcal{M}(\mathbf{z}_i), \mathcal{M}(\mathbf{z}_i'))$. Similarly, the hyperbolic supervised self-expertise loss is defined as:

$$\mathcal{L}_{\text{SSE}}^{\mathbb{H}} = \frac{1}{2}\left(\sum_{k=0}^{\lg K} \frac{\alpha_d(\mathcal{L}_{dis}^k|\frac{\mathbf{d}}{2^k}) + (1-\alpha_d)(\mathcal{L}_{ang}^k|\frac{\mathbf{d}}{2^k})}{2^k}\right), \tag{3}$$

where $\mathcal{L}_{dis}^k|\frac{\mathbf{d}}{2^k}$ and $\mathcal{L}_{ang}^k|\frac{\mathbf{d}}{2^k}$ denote the hyperbolic supervised distance-based and angle-based losses applied exclusively to the segment $\frac{\mathbf{d}}{2^k}$. The final training objective of Hyp-SelEx is formulated as:

$$\mathcal{L}_{rep}^{\mathbb{H}} = (1-\lambda_b^{\mathbb{H}})\mathcal{L}_{\text{USE}}^{\mathbb{H}} + \lambda_b^{\mathbb{H}}\mathcal{L}_{\text{SSE}}^{\mathbb{H}}. \tag{4}$$

B. More Quantitative Results

B.1. GCD With Unknown Category Numbers

In line with the majority of the literature [13, 16, 20, 21], our primary experiments presented in the main paper utilize the ground-truth category numbers. This section reports results based on estimated category numbers obtained from an off-the-shelf method [16], illustrating the performance of our approach when ground-truth category numbers are unavailable. For the CUB dataset, we estimate $K = 231$, while for Stanford-Cars, we estimate $K = 230$. In contrast, the actual ground-truth counts are $K = 200$ and $K = 196$, respectively. We compare our methods with SimGCD [21], μGCD [18], and GCD [16] in Table 2. Despite a discrepancy of approximately 15% between the ground-truth and estimated category numbers for both CUB [19] and Stanford-Cars [6], our hyperbolic methods exhibit only a marginal decline in performance.

Table 2. Results with the estimated number of categories, all methods use the DINO [2] pretrained weights.

	CUB [19]			Stanford-Cars [6]		
Method	All	Old	New	All	Old	New
GCD [16]	47.1	55.1	44.8	35.0	56.0	24.8
SimGCD [21]	61.5	66.4	59.1	49.1	65.1	41.3
μGCD [18]	62.0	60.3	62.8	56.3	66.8	51.1
SelEx [13]	72.0	72.3	71.9	58.7	75.3	50.8
Hyp-GCD	60.2	64.6	58.0	48.1	60.2	42.2
Hyp-SimGCD	64.7	66.6	63.8	60.3	73.5	53.9
Hyp-SelEx	**79.6**	**75.8**	**81.6**	**62.1**	**76.2**	**55.3**

Table 3. Experimental results using different embedding dimensions on Hyp-GCD with DINO [2] pre-trained backbone. Results on the CUB [19] and Stanford-Cars [6] datasets are reported.

	CUB [19]			Stanford-Cars [6]		
dimension	All	Old	New	All	Old	New
64	57.6	63.6	54.6	47.2	56.7	42.6
128	59.5	65.0	56.7	48.2	60.0	42.5
256	61.0	**67.0**	58.0	**50.8**	**60.9**	45.8
512	**61.2**	65.3	**59.1**	50.3	59.5	**45.9**

B.2. Embedding Dimension

In our framework, the parametric method Hyp-SimGCD employs the original 768-dimensional embeddings from the pre-trained ViT-B backbone. For the non-parametric methods, Hyp-GCD and Hyp-SelEx, we project the features from the pretrained backbone into a new spherical space using an MLP projection network, followed by an exponential mapping into hyperbolic space. In the baseline methods, GCD and SelEx, the final embedding dimension is set to $65,536$. However, our empirical findings indicate that a significantly lower dimension can yield satisfactory performance with our hyperbolic method, Hyp-GCD. As shown in Tab. 3, embeddings of 256 dimensions yield promising results for Hyp-GCD. This suggests that the intrinsic properties of hyperbolic space facilitate more expressive representations at lower dimensions (*e.g.*, 256 or 512), effectively capturing hierarchical structures and complex relationships among data points. For Hyp-SelEx, we have chosen a dimension of $8,092$, which is also significantly lower than that of the baseline methods.

Table 4. Comparison with recent GCD methods on Herbarium19 [15] and Oxford-Pet [10].

	Oxford-Pet [10]			Herbarium19 [15]		
Method	All	Old	New	All	Old	New
ORCA [1]	-	-	-	24.6	26.5	23.7
GCD [16]	80.2	85.1	77.6	35.4	51.0	27.0
XCon [5]	86.7	91.5	84.1	-	-	-
OpenCon [14]	-	-	-	39.3	58.9	28.6
DCCL [11]	88.1	88.2	88.0	-	-	-
SimGCD [21]	91.7	83.6	**96.0**	44.0	58.0	36.4
μGCD [18]	-	-	-	**45.8**	**61.9**	**37.2**
InfoSieve [12]	90.7	**95.2**	88.4	40.3	59.0	30.2
SelEx [13]	92.5	91.9	92.8	39.6	54.9	31.3
Hyp-GCD	86.7	85.5	87.4	38.6	43.1	36.2
Hyp-SimGCD	92.2	85.7	95.7	45.1	60.1	36.9
Hyp-SelEx	**92.7**	91.5	93.3	40.5	49.0	36.0

B.3. Results on Additional Datasets

To further evaluate the proposed method, we conduct assessments on two additional fine-grained datasets: Oxford-Pet[10] and Herbarium19[15]. The Oxford-Pet dataset poses a significant challenge due to its variety of cat and dog species, alongside limited data availability. In contrast, Herbarium19 is a botanical research dataset that encompasses a wide range of plant types, characterized by its long-tailed distribution and fine-grained categorization. The results of our experiments on these

two datasets are summarized in Tab. 4. Our Hyp-SelEx method achieves the highest accuracy across all categories in the Oxford-Pet dataset. Furthermore, on Herbarium19, Hyp-SelEx secures the second-best performance on all three evaluation metrics.

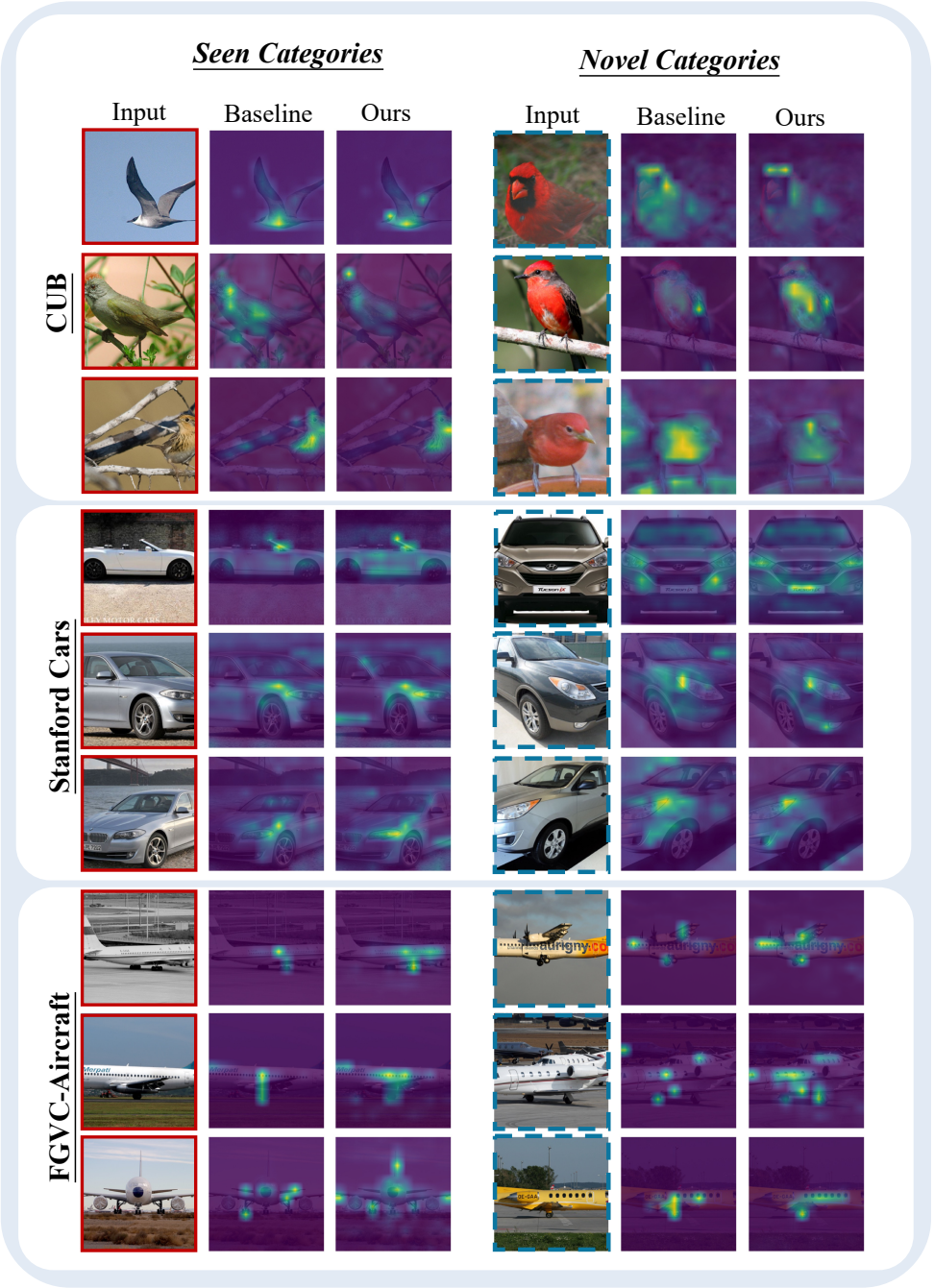

Figure 1. Visualization of attention maps of GCD [16] and our Hyp-GCD.

C. More Qualitative Results

Fig. 1 displays the attention maps of GCD [16] and Hyp-GCD, generated from the final transformer block of the DINO backbone [2]. These attention maps are applied across three fine-grained datasets within the SSB benchmark [17]. In this block, a *multi-head self-attention* layer utilizing 12 attention heads processes the input features, resulting in 12 attention maps at a resolution of 14×14. Following the methodology detailed in [2], we compute the mean value of these attention maps and subsequently upsample them to the original image resolution for visualization. The results indicate that our method significantly enhances focus on semantically relevant regions within the image, effectively capturing fine-grained details that are crucial for distinguishing between categories. In contrast, the baseline approach yields more diffuse and less targeted attention maps, often insufficiently highlighting critical areas, particularly concerning unseen categories. These findings emphasize the robustness and generalization capability of our method in identifying meaningful visual regions, even for novel categories, thereby demonstrating its superiority over the baseline approach.

References

[1] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In *ICLR*, 2022. 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. 1, 3, 5

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

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[5] Yixin Fei, Zhongkai Zhao, Siwei Yang, and Bingchen Zhao. Xcon: Learning with experts for fine-grained category discovery. In *BMVC*, 2022. 3

[6] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshop*, 2013. 1, 2, 3

[7] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

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

[9] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[10] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 1, 3

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

[12] Sarah Rastegar, Hazel Doughty, and Cees Snoek. Learn to categorize or categorize to learn? self-coding for generalized category discovery. In *NeurIPS*, 2023. 3

[13] Sarah Rastegar, Mohammadreza Salehi, Yuki M Asano, Hazel Doughty, and Cees G M Snoek. Selex: Self-expertise in fine-grained generalized category discovery. In *ECCV*, 2024. 1, 2, 3

[14] Yiyou Sun and Yixuan Li. Opencon: Open-world contrastive learning. *TMLR*, 2022. 3

[15] Kiat Chuan Tan, Yulong Liu, Barbara Ambrose, Melissa Tulig, and Serge Belongie. The herbarium challenge 2019 dataset. *arXiv preprint arXiv:1906.05372*, 2019. 1, 3

[16] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *CVPR*, 2022. 1, 2, 3, 4, 5

[17] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. The semantic shift benchmark. In *ICML workshop*, 2022. 5

[18] Sagar Vaze, Andrea Vedaldi, and Andrew Zisserman. No representation rules them all in category discovery. In *NeurIPS*, 2023. 2, 3

[19] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1, 2, 3

[20] Hongjun Wang, Sagar Vaze, and Kai Han. Sptnet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. In *ICLR*, 2024. 2

[21] Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *ICCV*, 2023. 1, 2, 3