Supplementary Material for "OmniGCD: Abstracting Generalized Category Discovery for Modality Agnosticism"

Jordan Shipard[1,2*] Arnold Wiliem[1,2] Kien Nguyen Thanh[1] Wei Xiang[3] Clinton Fookes[1]
[1]SAIVT, QUT, Australia [2]Shield AI, Australia [3]La Trobe University, Australia

In our supplementary material, we present the following sections in support of the main paper. We first provide qualitative examples showing GCDformer's ability to optimize the GCD latent spaces for k-means clustering in Section 1. We then provide further details on the setup for OmniGCD's zero-shot GCD experiments in Section 2. Next, in Section 3, we report ablation results for training GCDformer with higher data dimensionality than the 2D version used for the main results. In Section 4, we present results using additional encoders for the vision modality, specifically MobileNetV3 [2] and DINOv3 [10]. We also include standard deviation reporting for the results in Table 4 of the main paper in Section 5. Additionally, we provide a more detailed analysis of OmniGCD performance by inspecting the contributions of the dimension reduction method and GCDformer. Section 6 contains fine-tuning results for the datasets not presented in Table 5 of the main paper. Lastly, Section 7 contains additional results comparing OmniGCD performance across the considered dimension reduction methods.

1. GCDFormer Optimization of GCD Latent Spaces

The GCDformer in OmniGCD is trained exclusively on synthetic data to optimize the GCD latent space for clustering. In Figure 1, we present qualitative examples illustrating OmniGCD's ability to enhance GCD latent spaces. These examples comprise randomly selected synthetic GCD latent spaces from GCDformer's training. Overall, these visualizations demonstrate GCDformer's ability to optimize the input GCD latent spaces for improved clustering. Moreover, they highlight the adaptability of GCDformer. For instance, in latent spaces with already well-separated clusters, GCDformer tightens the clusters while preserving separation. In contrast, for spaces featuring many overlapping tight clusters, it increases inter-cluster separation. In all cases, GCDformer relies solely on labels from a random subset of points to guide the optimization.

*Work was done during PhD at SAIVT, QUT.
Emails: {jordan.shipard, arnold.wiliem}@shield.ai, {k.nguyenthanh, c.fookes}@qut.edu.au, w.xiang@latrobe.edu.au

2. Details regarding OmniGCDs Zero-shot setup

In this section, we provide additional details on the sizes of the labeled and unlabeled datasets used by OmniGCD for the main results in Table 4 of the main paper. As noted, at test time, OmniGCD's input comprises both the labeled and unlabeled subsets. In the zero-shot GCD setting, the labeled subset is constructed from training-split samples belonging to the labeled classes, while the unlabeled subset is the test set. During experimentation, we observed that OmniGCD's test-set performance varies slightly with the number of labeled samples per class drawn from the training data. For example, this variation is typically less than one percentage point in *All*, *Old*, and *New* accuracy for ± 5 samples per class. Given this variation, we performed hyperparameter sweeps to identify the optimal number of labeled samples per class for each dataset and encoder. These optimal values, used for all reported results, are listed in Table 1. We encourage future work to further examine the impact of the number of labeled samples per class, as a key desirable trait of a modality-agnostic GCD method is improved performance with additional samples per class.

3. OmniGCD Dimension Ablations

In the main paper, we chose 2D data for GCDformer training, as low dimensionality ensures tractability in generating training data. In this section, we present ablations on training with higher-dimensional data (specifically, 32, 64, and 128 dimensions). Due to the $O(N^2k)$ time complexity of t-SNE [3]—where N is the number of data points and k the number of dimensions—we conduct only limited ablations on the vision modality using DINOv1 [1]. The results in Table 2 show that OmniGCD's overall performance degrades as the input dimensionality increases. This validates our design choice of 2D data for GCDformer.

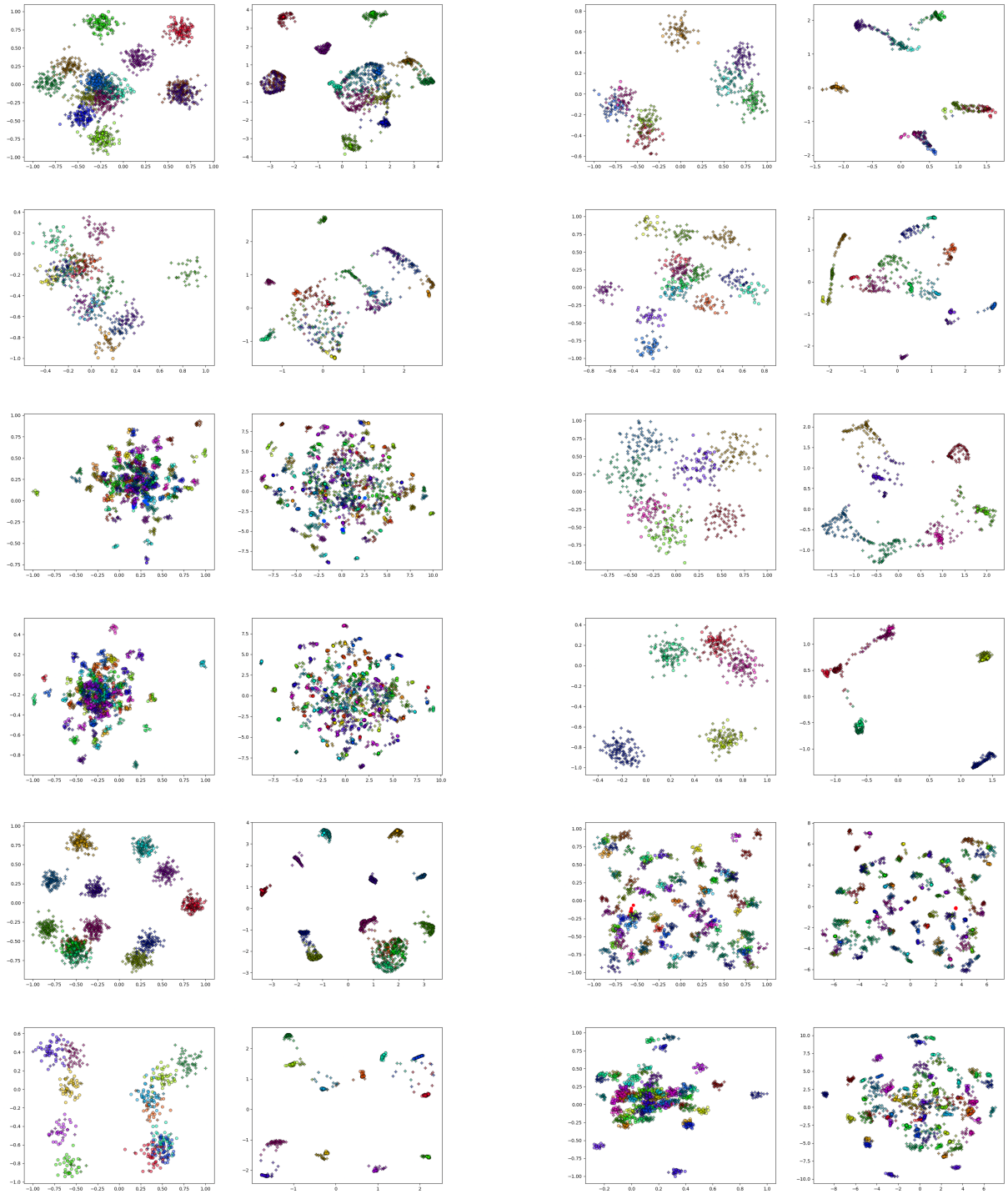

Figure 1. Qualitative examples of GCDformers ability to optimize the GCD latent spaces for k-means clustering. We show 10 examples of initial (left for each example) and opitmized GCD latent spaces (right for each example). These examples are randomly selected from GCDformers training. In these figures, circular points (○) represent data belonging to the labeled subset D_L, while data belonging to the unknown subnet D_U is represented by a plus sign (✚)

Dataset	Encoder	Labeled Data per Class	Total Labeled Data	Total Unlabeled Data
CIFAR10	DINOv1	90	450	10,000
	DINOv2	90	450	10,000
CIFAR100	DINOv1	150	12,000	10,000
	DINOv2	30	2,400	10,000
IN100	DINOv1	20	1,000	5,000
	DINOv2	10	500	5,000
CUB-200	DINOv1	25	2,500	5,794
	DINOv2	25	2,500	5,794
SCars	DINOv1	45	3,988	8,041
	DINOv2	35	3,390	8,041
Aircraft	DINOv1	20	1,000	3,333
	DINOv2	15	750	3,333
Herb-19	DINOv1	5	1,705	2,697
	DINOv2	20	6,506	2,697
BANKING	e5-Large	80	2,967	3,080
StackOverflow	e5-Large	40	400	1,000
CLINIC	e5-Large	10	750	3,000
VocalSet	MERT-95M	60	478	2,285
UrbanSound	MERT-95M	60	300	4,366
EuroSAT	DOFA-B	50	250	5,400
So2SAT	DOFA-B	20	140	4,838
RESIC45	DOFA-B	80	1,200	6,300
UC Merced	DOFA-B	50	446	420

Table 1. Number of labeled data per class, total labeled data and total unlabeled data used for optimal results of our GCD Transformer. These ratios were used for results presented in Table 4 of the main paper.

4. Additional Vision Encoders Ablations

Here, we present results using two additional pretrained vision encoders: MobileNetV3 [2] and DINOv3 [10]. These results are not merely to demonstrate that OmniGCD works with additional encoders, as the main results with multimodal encoders already establish this, but serve two specific purposes: (1) MobileNetV3 is a lightweight encoder suitable for edge devices, and we aim to demonstrate OmniGCD's performance with such an encoder; (2) DINOv3 is the most recent and highest-performing encoder in the DINO family, and we provide these results for future GCD studies that will likely adopt this encoder. As shown in Table 3, with MobileNetV3, OmniGCD achieves the best *All* accuracy on 5 of the 7 datasets, the best *Old* accuracy on 4 of the 7 datasets, and the best *New* accuracy on 3 of the 7 datasets. With DINOv3, it achieves the best *All* accuracy on all 7 datasets, the best *Old* accuracy on 3 of the 7 datasets, and the best *New* accuracy on 3 of the 7 datasets.

5. Further analysis of OmniGCD Zero-shot GCD performance

In this section, we provide a more granular analysis of OmniGCD's performance, along with standard deviations for the main results in Table 4 of the main paper. These results are presented in Table 5. For the granular analysis, we compare the individual contributions of t-SNE dimensionality reduction and GCDformer. Using neither component yields k-means clustering directly on the feature vectors, while using both yields the full OmniGCD method. We omit the variant using GCDformer without t-SNE, as it would require training separate GCDformer models per modality. Overall, combining t-SNE and GCDformer improves clustering performance over t-SNE alone in 80% of measured metrics across all datasets. The sole exception is *New* accuracy on the Aircraft [6] dataset with DINOv2 [8], where the combination worsens performance. In 5 instances, GCDformer fails to further improve clustering over t-SNE alone, and in 6 instances, it degrades performance. Notably, we observe the counterintuitive yet consistent improvements in GCD performance from t-SNE alone, despite the information loss from reducing feature vectors to 2D. We would expect such dimensionality reduction to impair clustering, but it does not. This suggests that dimensionality reduction is a valuable tool for GCD methods. It represents an intriguing discovery, hinting at an unintentional alignment between dimensionality reduction and GCD tasks. We encourage future work to explore this further, as exploiting such alignment could enhance GCD performance in subsequent methods.

Num Dimensions	CIFAR-10			CIFAR-100			ImageNet-100			CUB-200			Stanford Cars			Aircraft			Herbarium-19		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
2	**90.7**	**97.0**	91.6	**60.0**	**89.9**	**74.6**	**81.1**	**91.0**	**74.8**	**44.5**	**66.0**	33.9	**12.6**	**19.6**	**16.9**	**18.9**	**14.3**	**13.9**	**30.8**	**48.6**	**50.0**
32	87.2	94.6	91.0	58.5	70.8	73.8	78.5	90.0	74.2	35.9	58.0	**45.4**	**12.6**	13.7	16.3	13.2	9.0	10.3	26.8	37.1	45.0
64	51.7	67.6	27.9	56.9	71.5	69.4	61.8	78.0	46.4	28.9	43.3	26.8	12.3	12.8	14.9	12.9	13.1	7.9	27.2	42.9	**50.0**
128	74.7	95.8	**91.7**	39.7	82.1	57.1	65.1	86.0	43.8	23.9	54.0	45.1	8.1	4.8	6.1	12.3	11.3	9.1	22.8	42.9	**50.0**

Table 2. Ablation of zero-shot GCD performance for OmniGCD using the DINOv1 (ViT-B/16) [1] vision encoder. Features from DINOv1 are projected to varying t-SNE embedding dimensions (2, 32, 64, 128) across the vision modality datasets. We **bold** the best results for each comparison.

		CIFAR-10			CIFAR-100			ImageNet-100			CUB-200			Stanford Cars			Aircraft			Herbarium-19		
		All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
MobileNetV3	K-means	66.4	74.7	71.0	35.6	16.5	41.2	69.5	59.9	64.7	45.5	34.5	45.2	14.2	34.9	8.8	15.3	14.8	16.2	25.8	12.9	**18.3**
	GCD (w/o FT)	71.2	**89.0**	**82.0**	**48.4**	**48.5**	26.5	56.6	46.0	76.0	44.4	50.0	38.3	**17.8**	**64.4**	**11.9**	16.9	6.0	**19.7**	26.4	**14.3**	11.1
	OmniGCD	**77.5**	72.5	70.6	40.2	23.5	**44.0**	**77.1**	**61.0**	**84.0**	**54.2**	**53.0**	**53.7**	16.4	23.0	7.4	**17.2**	**15.8**	18.8	**28.3**	**14.3**	16.7
DINOv3	K-means	86.1	67.5	95.6	66.6	49.7	96.3	78.2	27.8	83.8	71.3	76.7	58.8	58.0	60.6	72.6	39.0	38.8	59.0	30.2	**34.9**	32.3
	GCD (w/o FT)	80.5	76.6	95.5	70.6	**72.5**	**97.0**	67.0	36.0	**88.0**	63.6	**95.0**	29.9	56.4	45.0	46.7	36.5	**41.8**	**62.7**	33.5	33.3	**50.0**
	OmniGCD	**97.2**	**95.1**	**98.8**	**74.7**	40.1	93.4	**88.2**	**38.0**	85.0	**79.3**	84.7	**86.7**	**68.1**	**83.0**	**89.1**	**44.0**	36.1	59.7	**36.0**	33.3	25.0

Table 3. Additional Zero-shot GCD results for the vision modality using the MobileNetV3 (MBV3-Large) [2] and DINOv3 (ViT-B/16) [10] encoders. We **bold** the best results for each comparison.

	CIFAR-10			CIFAR-100			ImageNet-100			CUB-200			Stanford Cars			Aircraft			Herbarium-19		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
GCD	81.8	86.2	76.9	69.0	77.4	62.0	73.5	**92.6**	63.9	51.3	56.6	48.7	39.0	57.6	29.9	45.0	41.1	46.9	**35.4**	51.0	27.0
OmniGCD (w/o FT)	90.7	**97.0**	91.6	60.0	**89.9**	**74.6**	**81.1**	91.0	74.8	44.5	66.0	33.9	12.6	19.6	16.9	18.9	14.3	13.9	30.8	48.6	**50.0**
OmniGCD (w FT)	**96.1**	96.2	**96.1**	**71.7**	76.8	66.0	76.2	91.6	**94.8**	**59.8**	**71.0**	**90.6**	**49.1**	**63.7**	**48.1**	**47.5**	**71.21**	**59.4**	34.4	**79.2**	39.6

Table 4. Results on the standard GCD setting compared to the original GCD method which fine-tunes its encoder [13]. We fine-tune the vision encoder using the same way as the original GCD method [13]. We **bold** the best results for each comparison.

	t-SNE	GCDformer	CIFAR-10			CIFAR-100			ImageNet-100			CUB-200			Stanford Cars			Aircraft			Herbarium-19		
			All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
DINOv1 [1]	✗	✗	77.8 ± 6.6	88.1 ± 2.2	60.3 ± 38.6	52.5 ± 0.6	68.9 ± 12.1	66.3 ± 5.8	73.5 ± 1.3	87.6 ± 2.5	69.9 ± 17.4	35.8 ± 1.0	54.7 ± 9.3	39.8 ± 9.4	10.7 ± 0.2	14.5 ± 2.5	15.9 ± 1.7	14.8 ± 0.7	10.9 ± 4.6	10.8 ± 5.2	24.9 ± 0.4	42.9 ± 10.7	22.5 ± 11.5
	✓	✗	83.6 ± 6.3	91.5 ± 10.1	85.6 ± 13.1	57.7 ± 0.8	79.1 ± 8.4	74.5 ± 2.9	78.4 ± 1.2	89.8 ± 2.4	**74.8** ± 0.4	44.2 ± 0.5	64.0 ± 3.3	**38.6** ± 5.6	**12.6** ± 0.1	10.8 ± 2.8	13.0 ± 0.5	18.2 ± 0.2	11.3 ± 1.5	**17.6** ± 7.2	30.2 ± 0.2	42.9 ± 0.0	40.0 ± 12.2
	✓	✓	**90.7** ± 3.9	**97.0** ± 3.1	**91.6** ± 0.4	**60.0** ± 0.8	**89.9** ± 7.4	**74.6** ± 16.9	**81.1** ± 1.5	**91.0** ± 0.0	**74.8** ± 0.5	**44.5** ± 1.5	**66.0** ± 2.7	33.9 ± 6.6	**12.6** ± 0.2	**19.6** ± 13.6	**16.9** ± 3.6	**18.9** ± 0.8	**14.3** ± 3.3	13.9 ± 6.4	**30.8** ± 0.2	**48.6** ± 5.7	**50.0** ± 0.0
DINOv2 [8]	✗	✗	82.1 ± 8.5	62.6 ± 25.4	95.0 ± 4.3	69.1 ± 1.8	44.2 ± 20.4	51.3 ± 6.4	79.9 ± 1.3	87.7 ± 1.7	78.4 ± 0.7	70.3 ± 0.9	95.8 ± 4.9	75.4 ±	27.3 ± 0.6	16.8 ± 6.7	28.7 ± 3.8	19.6 ± 0.3	12.7 ± 4.9	**22.5** ± 3.9	29.0 ± 0.5	40.8 ± 5.1	23.2 ± 3.3
	✓	✗	89.9 ± 4.6	93.3 ± 8.6	91.0 ± 7.8	75.2 ± 0.7	**48.5** ± 8.6	47.0 ± 8.0	86.8 ± 1.3	84.4 ± 18.2	**90.0** ± 0.0	77.5 ± 0.3	99.3 ± 1.3	79.29 ± 21.0	33.3 ± 0.4	23.2 ± 3.4	38.4 ± 6.7	**21.2** ± 0.2	15.8 ± 1.2	18.2 ± 6.9	33.5 ± 0.2	**60.0** ± 3.1	**35.5** ± 4.1
	✓	✓	**96.9** ± 3.5	**96.9** ± 0.7	**95.6** ± 0.4	**78.1** ± 1.3	47.9 ± 4.5	**56.8** ± 9.6	**88.7** ± 0.4	**94.0** ± 0.0	**90.0** ± 0.0	**79.8** ± 0.9	**100.0** ± 0.0	**96.4** ± 0.0	**33.4** ± 0.6	**24.2** ± 2.8	**43.1** ± 1.1	**21.2** ± 0.1	**17.0** ± 2.2	11.6 ± 7.4	**34.8** ± 0.5	**60.0** ± 1.5	25.8 ± 9.7

		BANKING			StackOverflow			CLINIC			VocalSet			UrbanSound			EuroSAT			So2SAT		
		All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
✗	✗	57.7 ± 1.5	64.4 ± 14.3	43.4 ± 8.2	72.9 ± 3.7	68.3 ± 11.8	45.2 ± 9.3	69.5 ± 1.6	85.0 ± 13.4	81.7 ± 16.0	22.1 ± 1.5	15.0 ± 2.7	25.0 ± 6.0	39.3 ± 2.0	37.6 ± 10.6	37.7 ± 11.7	53.2 ± 2.4	38.4 ± 6.2	62.4 ± 7.1	27.3 ± 1.1	32.5 ± 2.3	65.1 ± 1.1
✓	✗	65.2 ± 1.2	41.3 ± 6.0	**51.1** ± 10.0	84.8 ± 3.1	**90.2** ± 6.6	**85.4** ± 5.4	81.4 ± 0.8	85.0 ± 6.4	72.7 ± 11.3	22.1 ± 1.3	14.7 ± 1.3	32.7 ± 7.0	44.8 ± 2.4	53.8 ± 5.8	52.5 ± 10.3	66.0 ± 2.5	71.1 ± 2.3	61.0 ± 14.5	31.1 ± 1.8	**39.5** ± 3.6	48.2 ± 5.9
✓	✓	**66.4** ± 2.2	**77.1** ± 34.8	40.8 ± 17.8	**86.8** ± 2.0	78.0 ± 15.2	73.8 ± 12.3	**82.1** ± 1.6	**96.0** ± 5.6	**90.0** ± 12.1	**25.6** ± 2.3	**17.5** ± 4.7	**49.2** ± 5.8	**45.5** ± 1.0	**56.7** ± 9.7	**63.9** ± 5.3	**68.3** ± 6.3	**75.5** ± 0.8	**69.5** ± 19.0	**31.7** ± 0.5	38.4 ± 2.8	**58.4** ± 2.9

		RESIC45			UC Merced		
		All	Old	New	All	Old	New
✗	✗	42.8 ± 2.2	37.1 ± 7.7	60.8 ± 8.6	63.6 ± 3.0	50.6 ± 10.4	70.8 ± 12.9
✓	✗	55.6 ± 2.5	60.4 ± 19.4	71.3 ± 7.7	64.1 ± 4.6	52.4 ± 12.7	74.1 ± 12.2
✓	✓	**58.5** ± 0.9	**73.0** ± 13.4	**74.4** ± 1.0	**75.8** ± 4.8	**67.7** ± 14.7	**95.7** ± 0.4

Table 5. Results comparing the individual contributions of the t-SNE [12] dimension reduction method and GCDformer for optimizing the GCD latent spaces. The use of each component is denoted as a ✓ if it is used or a ✗ if it is not used. We also report the standard deviation for these results. The usage of both t-SNE and GCDformer is the full OmniGCD method, as such these results are the same as Table 4 of the main paper. We **bold** the best results for each comparison.

		CIFAR-10			CIFAR-100			ImageNet-100			CUB-200			Stanford Car			Aircraft			Herbarium-19		
		All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
DINOv1 [1]	PCA	47.2	66.6	45.3	10.7	13.2	10.5	16.6	9.4	4.8	10.6	11.3	20.3	5.9	5.5	5.8	9.4	7.8	5.8	25.9	28.6	45.0
	UMAP	**91.4**	96.4	**94.0**	57.6	82.6	74.3	77.2	89.0	77.0	41.3	60.0	25.1	10.4	5.5	11.9	16.2	11.9	11.2	27.7	31.4	45.0
	t-SNE	90.7	**97.0**	91.6	**60.0**	**89.9**	**74.6**	**81.1**	**91.0**	**74.8**	**44.5**	**66.0**	**33.9**	**12.6**	**19.6**	**16.9**	**18.9**	**14.3**	**13.9**	**30.8**	**48.6**	**50.0**
DINOv2 [8]	PCA	70.3	64.5	84.0	12.7	9.0	2.1	20.6	7.8	6.8	16.9	27.7	12.1	13.1	21.2	16.7	16.0	22.7	**12.2**	26.1	13.9	9.7
	UMAP	**97.1**	**98.0**	79.1	73.1	**52.8**	62.4	84.8	**97.0**	**91.0**	76.7	**100.0**	**98.2**	30.1	**24.4**	43.0	18.1	12.7	3.9	30.9	41.5	15.5
	t-SNE	96.9	96.9	**95.6**	**78.1**	47.9	**56.8**	**88.7**	94.0	90.0	**79.8**	**100.0**	96.4	**33.4**	24.2	**43.1**	**21.2**	**17.0**	11.6	**34.8**	**60.0**	**25.8**

		BANKING			StackOverflow			CLINIC			VocalSet			UrbanSound			EuroSAT			So2SAT [15]		
		All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
	PCA	17.5	16.0	25.6	35.3	34.8	43.2	13.8	8.7	11.0	19.8	14.2	29.3	32.3	36.8	44.6	37.7	46.5	56.6	26.3	28.0	61.1
	UMAP	65.4	**86.0**	22.8	**87.6**	75.0	**74.2**	**82.9**	45.2	68.1	21.1	14.4	29.4	41.2	21.8	52.6	55.1	68.7	46.7	30.8	31.5	37.9
	t-SNE	**66.4**	77.1	**40.8**	86.8	**78.0**	73.8	82.1	**96.0**	**90.0**	**25.6**	**17.5**	**49.2**	**45.5**	**56.7**	**63.9**	**68.3**	**75.5**	**69.5**	**31.7**	**38.4**	**58.4**

		RESIC45			UC Merced		
		All	Old	New	All	Old	New
	PCA	15.7	16.8	30.8	37.1	55.9	40.0
	UMAP	56.1	67.8	69.3	73.0	**76.5**	74.5
	t-SNE	**58.5**	**73.0**	**74.4**	**75.8**	67.7	**95.7**

Table 6. Comparing OmniGCDs performance using PCA [9], UMAP [7] and t-SNE [12] across all datasets and modalities. t-SNE achieves the best performance overall. We **bold** the best results for each comparison.

6. Additional Results for Vision Encoder Fine-tuning

Here, we provide additional results for Table 5 in the main paper, focusing on OmniGCD's performance with a fine-tuned DINOv1 [1] encoder. As noted in the main paper, this vision encoder was fine-tuned using GCD's [13] supervised and self-supervised contrastive training methods. The results in Table 4 show that fine-tuning significantly improves performance on datasets where the base DINOv1 encoder performs poorly (CUB-200 [14], Stanford Cars [4], and Aircraft [6]). However, these gains are not consistent across all datasets. While *All* accuracy improves on CIFAR-10 [5], CIFAR-100 [5], and Herbarium-19 [11], fine-tuning yields lower *All* accuracy on ImageNet-100. Although, for ImageNet-100, it does improve *Old* and *New* accuracy. Overall, these findings indicate that fine-tuning may be necessary when encoders fail to accurately encode the target data; future methods should explore how to best integrate the benefits of fine-tuning with modality-agnostic GCD.

7. Dimension Reduction Options: Additional Results

In this section, we provide additional results for Table 8 in the main paper, comparing OmniGCD's performance across PCA [9], UMAP [7], and t-SNE [12] as dimensionality reduction methods. The results in Table 6 show that t-SNE yields the best performance on 54 of the 69 metrics measured across all datasets and modalities. Notably, UMAP ranks second, with PCA performing significantly worse overall. These results further validate our choice of t-SNE for OmniGCD.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 1, 4, 5, 6

[2] Andrew G. Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 1, 3, 4

[3] Sangwon Jung, Tristan Dagobert, Jean-Michel Morel, and Gabriele Facciolo. A review of t-sne. *Image Processing On Line*, 14:250–270, 2024. 1

[4] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 554–561, 2013. 6

[5] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images, 2009. Technical report, University of Toronto. 6

[6] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft, 2013. 3, 6

[7] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *J. Open Source Softw.*, 3(29):861, 2018. 5, 6

[8] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3, 5

[9] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572, 1901. 5, 6

[10] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 1, 3, 4

[11] Kiat Chuan Tan, Yulong Liu, Barbara Ambrose, Melissa Tulig, and Serge Belongie. The herbarium challenge 2019 dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 6

[12] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *J. Mach. Learn. Res.*, 9:2579–2605, 2008. 5, 6

[13] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 4, 6

[14] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 6

[15] Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Jian Kang, Lichao Mou, Hossein Bagheri, Matthias Haberle, Yuansheng Hua, Rong Huang, Lloyd Hughes, Hao Li, Yao Sun, Guichen Zhang, Shiyao Han, Michael Schmitt, and Yuanyuan Wang. So2sat lcz42: A benchmark data set for the classification of global local climate zones [software and data sets]. *IEEE Geosci. Remote Sens. Mag.*, 8(3):76–89, 2020. 5