Learning for Transductive Threshold Calibration in Open-World Recognition


Qin Zhang


Dongsheng An


Tianjun Xiao


Tong He


Qingming Tang


Ying Nian Wu

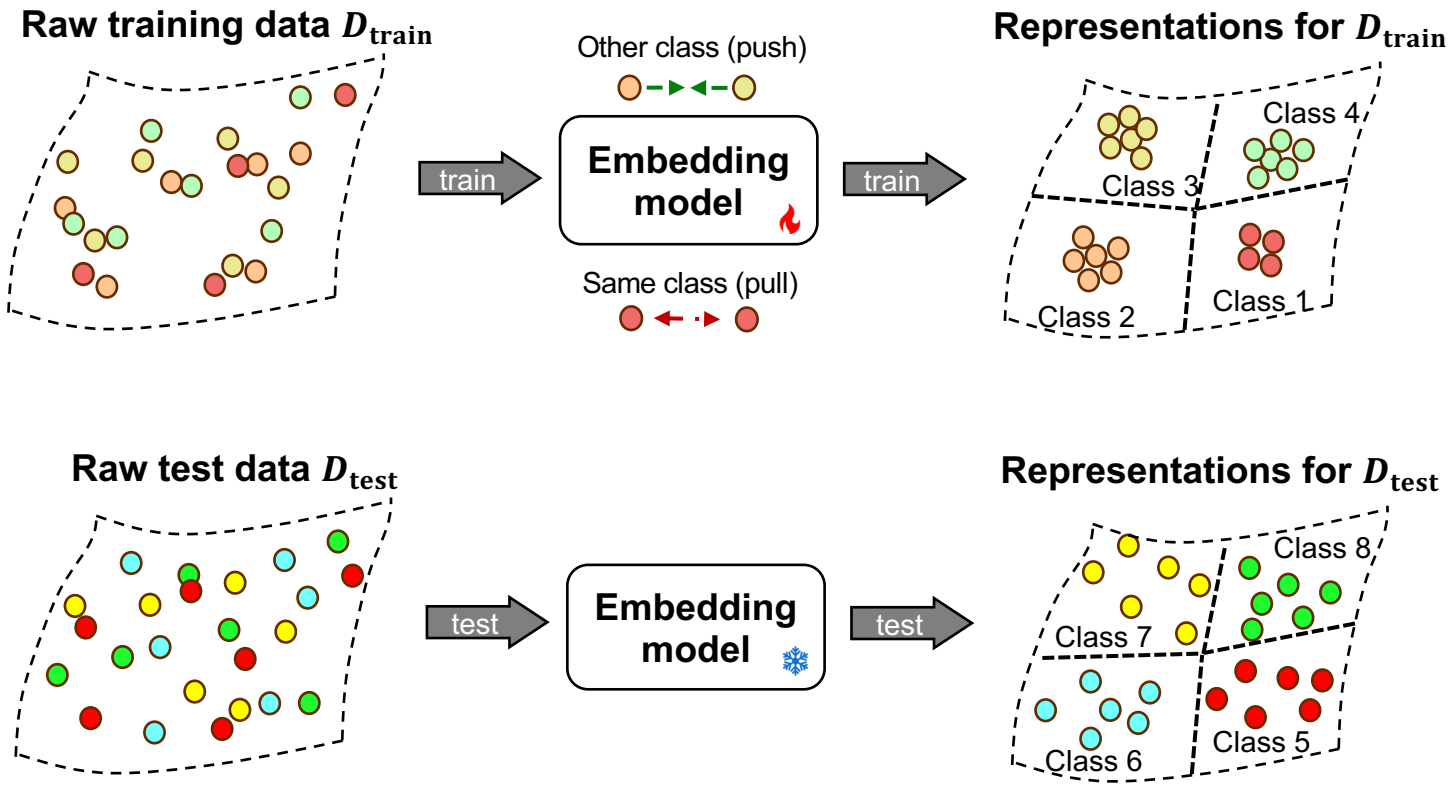

Joseph Tighe


Yifan Xing

amazon Rekognition

Open-World Recognition (OWR)

- Modern **Open-World Recognition (OWR)** systems typically use Deep Metric Learning to learn to transform raw data into vectorized representations, where distances between the representations reflect semantic similarities. During testing, the learned model is applied to unseen open-world classes (not encountered during training), with the expectation that similar items still remain close while dissimilar ones will be kept apart.



A few notes:

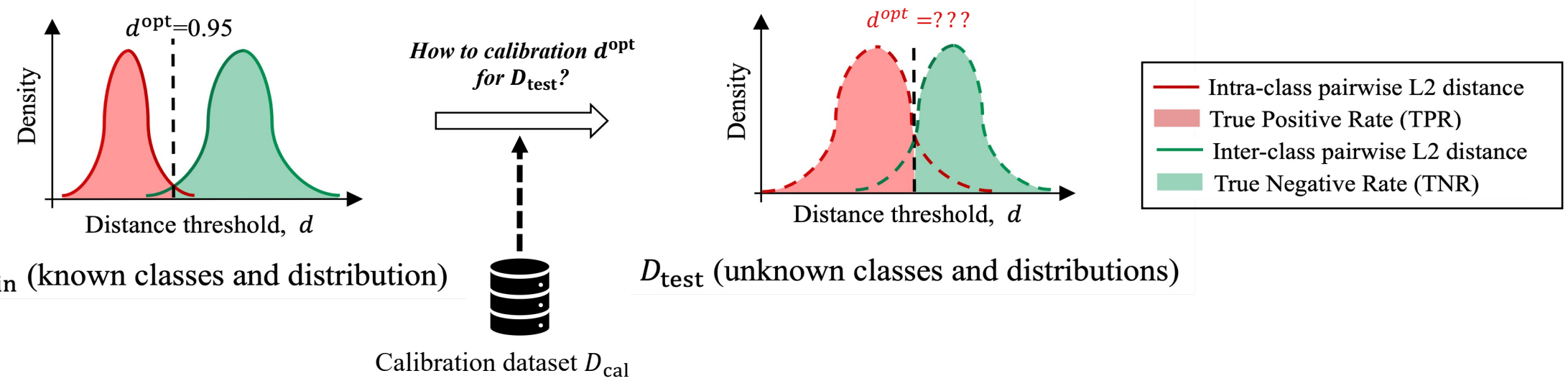
- The classes in D_{train} and D_{test} are disjoint.
- Even for the closed-world classes in D_{train}, the learned embedding model can exhibit significant variation in intra-class compactness and inter-class separations.
- When applied to the open-world classes in D_{test}, this variation in representation structures tends to become more severe.

Open-world Threshold Calibration

- **Problem Definition**: Open-world calibration amounts to choosing appropriate thresholds of open-world classes (e.g., wolf, sheep, mouse) for an embedding model trained on closed-set classes (dog, cat, bird) to balance the trade-off between TPR and TNR, considering potential distribution shifts and uncertainties in the open world.

- ***Mathematical formulations*** Let d and α be the distance threshold and minimum requirement for $\mathrm{TPR}_{\mathrm{test}}$, we formulate it as a constrained optimization problem:

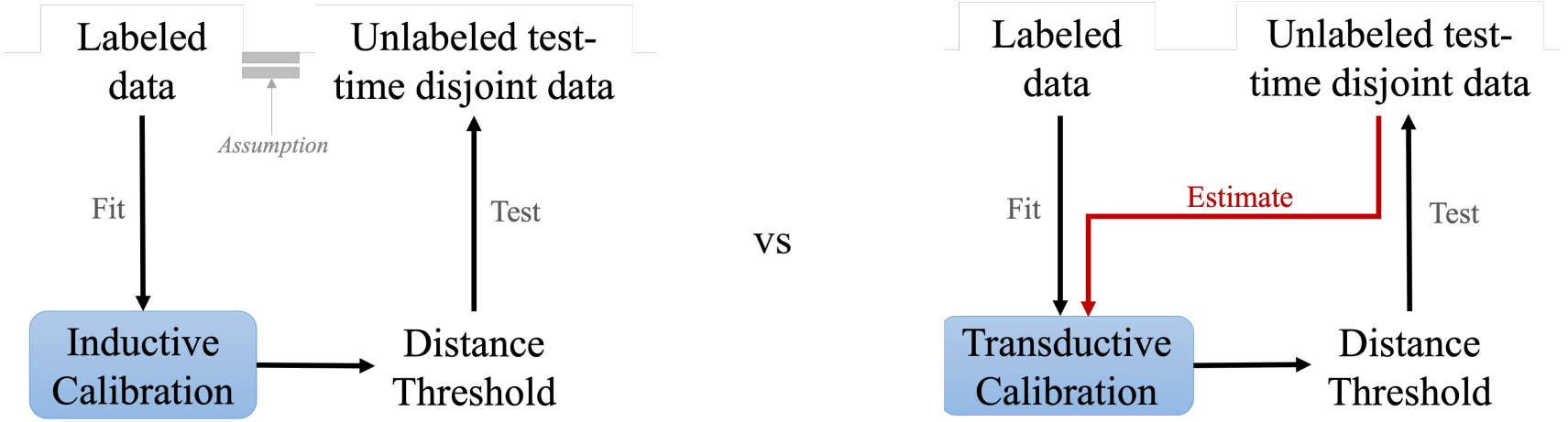
$$\underset{d}{\text{maximize}}\, \mathrm{TNR}_{\mathrm{test}}, \quad \text{subject to } \mathrm{TPR}_{\mathrm{test}}(d) \geq \alpha$$



Limitations of Existing Methods

- ***Problems of existing methods:*** Traditional calibration methods like *Platt Scaling and Isotonic Regression* are inductive, relying on a calibration dataset to learn general calibration rules under the assumption of identically distributed data. This assumption often fails in open-world scenarios, where the test distribution is unknown and dynamic, and the calibration dataset may not accurately represent the test data.

- ***Our remedial solution*** Transductive Threshold Calibration (TTC) .



(a) (Traditional) Inductive Calibration

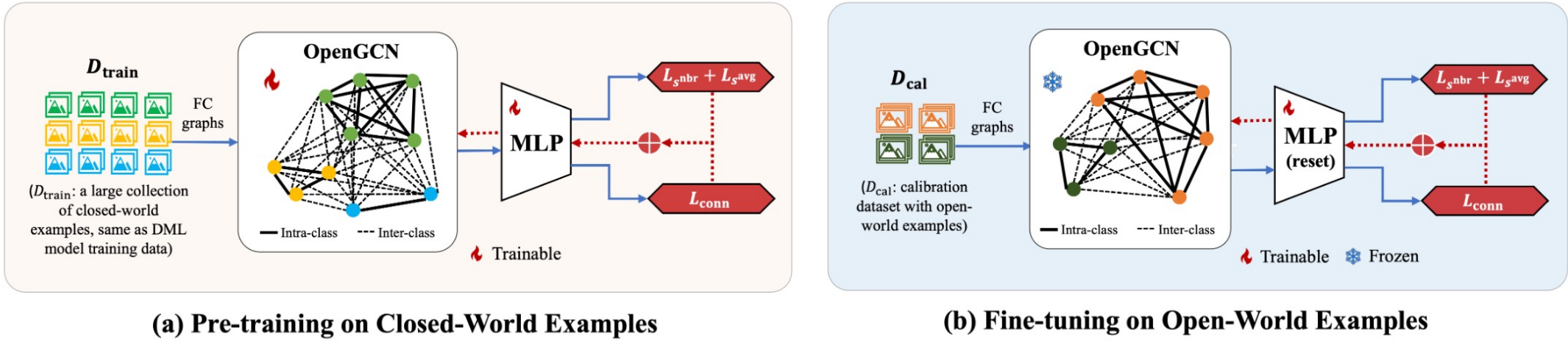
(b) (Our) Transductive Calibration

OpenGCN: Learning for Transductive Calibration

- **Main Design Concepts:**
  - Incorporate TTC by directly predicting the connectivity p_{ij} between data pairs using a Graph Neural Network (GNN), which can be used for predicting TPR and TNR for the test data at each distance threshold.

$$\hat{\mathrm{TPR}}_{\mathrm{test}}(d) = \sum_{i,j \in D_{\mathrm{test}}} 1_{p_{ij} > \tau} \cdot 1_{d_{ij} < d} \Big/ \sum_{i,j \in D_{\mathrm{test}}} 1_{p_{ij} > \tau} \qquad \hat{\mathrm{TNR}}_{\mathrm{test}}(d) = \sum_{i,j \in D_{\mathrm{test}}} 1_{p_{ij} \leq \tau} \cdot 1_{d_{ij} > d} \Big/ \sum_{i,j \in D_{\mathrm{test}}} 1_{p_{ij} \leq \tau}$$

  - Utilize both training data (for the OWR embedding model) and calibration data for calibration purpose, maximizing the rich information contained within the closed-set data.

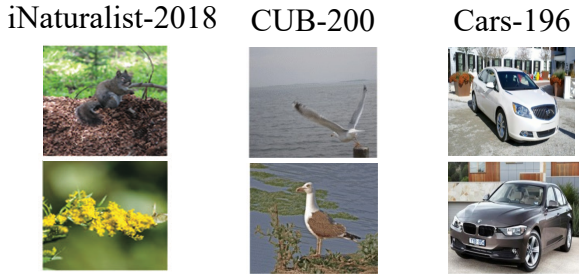


(a) Pre-training on Closed-World Examples

(b) Fine-tuning on Open-World Examples

Experiment Results – SameDist Scenario

- **Experiment settings:** SameDist – identical training and testing distributions.
- **Datasets:** iNaturalist-2018, CUB-200, Cars-196.
- **Evaluation metrics:** Absolute Error in TPR or TNR. We evaluate at multiple target values (TPR=80%, 90% and TNR=80%, 90%) to provide a comprehensive assessment.



$$\mathrm{AE}_{\mathrm{TPR}} = |\mathrm{TPR}(\hat{d}^{\mathrm{opt}}) - \mathrm{TPR}_{\mathrm{target}}| \qquad \mathrm{AE}_{\mathrm{TNR}} = |\mathrm{TNR}(\hat{d}^{\mathrm{opt}}) - \mathrm{TNR}_{\mathrm{target}}|$$

Table 1. Evaluation in the SameDist scenario using pointwise metrics of $\mathrm{AE}_{\mathrm{TPR}}$ (optimize for TPR) and $\mathrm{AE}_{\mathrm{TNR}}$ (optimize for TNR). The smaller the metric, the better. For each dataset, the best and second best results are marked in Red and Blue, respectively. Shading in the Table: Gray for posthoc calibration baselines, Cyan for clustering baselines, and Blue for our OpenGCN method. *Best viewed in color.*

Method	Optimize for TPR=80%			Optimize for TPR=90%			Optimize for TNR=80%			Optimize for TNR=90%			Rank
	Cars	CUB	Inat	Cars	CUB	Inat	Cars	CUB	Inat	Cars	CUB	Inat	
Platt scaling [37]	1.35%	5.10%	6.08%	0.44%	2.63%	4.63%	2.83%	2.02%	7.54%	2.93%	6.49%	0.92%	6
Beta calibration [24]	1.13%	5.16%	5.51%	**0.02%**	2.91%	3.26%	2.94%	1.41%	7.57%	2.78%	6.43%	0.93%	5
Isotonic regression [54]	0.82%	5.28%	4.53%	0.90%	2.56%	3.54%	1.94%	1.00%	5.78%	1.26%	4.65%	0.65%	3
Histogram Calibration [53]	0.82%	5.28%	4.53%	0.90%	2.56%	3.54%	1.94%	1.00%	5.78%	1.26%	4.65%	0.65%	4
DBSCAN [13]	43.11%	18.87%	**0.45%**	34.57%	9.18%	**1.85%**	4.09%	13.77%	12.90%	1.60%	9.32%	9.32%	7
Hi-LANDER [49]	3.44%	1.36%	10.54%	2.02%	**0.93%**	7.00%	**0.06%**	0.38%	2.35%	**0.10%**	2.20%	0.21%	2
OpenGCN (ours)	**0.33%**	**0.74%**	1.59%	0.72%	1.41%	2.37%	0.61%	**0.09%**	**0.74%**	0.58%	**0.72%**	**0.10%**	**1**

Experiment Results – ShiftDist Scenario

- **Experiment settings:** ShiftDist – slightly shifted distributions between training and calibration.
- **Evaluation metrics:** The combined Mean Absolute Error for both TPR and TNR averaged over $\mathrm{d} \in [0,2]$.

$$\mathrm{MAE}_{\mathrm{comb}} = \frac{1}{2}\int_0^2 \left(|\hat{\mathrm{TPR}}(\mathrm{d}) - \mathrm{TPR}(\mathrm{d})| + |\hat{\mathrm{TNR}}(\mathrm{d}) - \mathrm{TNR}(\mathrm{d})|\right) d\mathrm{d}$$

Table 2. Evaluation on the Cars-196 dataset in the ShiftDist scenario across 13 common corruption and perturbation types using combined global error metric of $\mathrm{MAE}_{\mathrm{comb}}$. The best results are marked in Red.

Method	Noise			Blur			Weather			Digital				Rank
	Gauss	Shot	Impulse	Defocus	Motion	Zoom	Snow	Fog	Bright	Contrast	Elastic	Pixel	JPEG	
Platt scaling	2.95e-2	2.99e-2	3.41e-2	2.66e-2	2.62e-2	5.02e-2	4.24e-2	4.37e-2	2.16e-2	4.61e-2	2.16e-2	2.24e-2	2.03e-2	4
Beta calibration	2.94e-2	2.97e-2	3.41e-2	2.67e-2	2.69e-2	5.06e-2	4.32e-2	4.37e-2	2.18e-2	4.61e-2	2.20e-2	2.23e-2	2.02e-2	5
Isotonic regression	2.88e-2	2.85e-2	3.38e-2	2.37e-2	2.31e-2	4.85e-2	4.07e-2	4.34e-2	1.83e-2	4.59e-2	1.85e-2	2.03e-2	1.80e-2	2
Histogram calibration	2.88e-2	2.85e-2	3.38e-2	2.37e-2	2.31e-2	4.85e-2	4.07e-2	4.34e-2	1.83e-2	4.59e-2	1.85e-2	2.03e-2	1.80e-2	3
DBSCAN	4.96e-2	6.02e-2	7.79e-2	9.81e-2	1.13e-1	1.22e-1	1.19e-1	4.02e-2	9.27e-2	4.53e-2	1.09e-1	1.04e-1	8.21e-2	7
Hi-LANDER	7.65e-2	6.30e-2	6.59e-2	3.98e-2	5.33e-2	4.48e-2	5.94e-2	7.16e-2	5.09e-2	9.45e-2	4.42e-2	9.48e-2	6.91e-2	6
OpenGCN (ours)	**1.33e-2**	**5.87e-3**	**1.66e-2**	**1.50e-2**	**1.71e-2**	**3.92e-2**	**1.42e-2**	**7.32e-3**	**6.73e-3**	**7.08e-3**	**5.34e-3**	**1.15e-2**	**1.68e-2**	**1**
Imp. over top baseline ↑	**53.82%**	**79.40%**	**50.89%**	**36.71%**	**25.97%**	**12.50%**	**65.11%**	**81.79%**	**63.22%**	**84.37%**	**71.14%**	**43.35%**	**6.67%**	**55.03%** (avg.)

Original | Gaussian noise | Shot noise | Impulse noise | Defocus blur | Motion blur | Zoom blur | Snow | Fog | Brightness | Contrast | Elastic transform | Pixelate | JPEG

Experiment Results – DiffDist Scenario

- **Experiment settings:** DiffDist – significantly different distributions between training and calibration data. In particular, we consider the following:
  - *Long-tailed calibration*: We divide iNaturalist's test classes into two sets based on cluster size, each containing the same number of images. For calibration, we use the head set (with a more images per class) as the calibration data and the tail set (with fewer images per class) for testing.
  - *Out-of-domain calibration*:
    - For Cars, we transform its test partition into sketches, while leaving the training and calibration partitions untouched.
    - We also consider cross-dataset calibration, where the OpenGCN model is pretrained and fine-tuned on iNaturalist (general natural species images) but tested on CUB (bird images).
- **Evaluation metrics:** The combined Mean Absolute Error.

Table 3. Evaluation in the DiffDist scenario using the global error metric MAE_{comb}. The best results are highlighted in Red.

Method	Cars: Sketch	CUB: Cross-dataset	iNat: Longtail
Platt scaling	1.08e-1	1.15e-1	2.09e-2
Beta calibration	1.08e-1	1.15e-1	2.12e-2
Isotonic regression	1.08e-1	1.15e-1	2.11e-2
Histogram Calibration	1.08e-1	1.15e-1	2.11e-2
DBSCAN	5.16e-2	1.60e-1	7.21e-2
Hi-LANDER	6.67e-2	1.30e-1	6.26e-2
OpenGCN (ours)	**3.54e-2**	**1.42e-2**	**1.82e-2**
Imp. over top baseline ↑	**31.40%**	**87.65%**	**12.92%**

Takeaways for “Learning for Transductive Threshold Calibration in Open-World Recognition”

amazon Rekognition

- We formally define the open-world threshold calibration problem for DML-based open-world visual recognition systems, identifying key challenges associated with the task.
- We introduce OpenGCN, a meta learning framework that enables transductive threshold calibration via a GNN. Importantly, OpenGCN does not rely on the assumption of matching distance distributions between the calibration dataset and the test dataset.
- The evaluation results underscore OpenGCN’s effectiveness across different distance distribution patterns between the calibration dataset and test dataset, highlighting its practical applicability for threshold calibration in DML-based open-world recognition.
- (Limitations) Compared to traditional calibration methods, OpenGCN is less efficient and more susceptible to over-parameterization. Furthermore, OpenGCN is not a calibration-data-free method as it still requires some calibration data in addition to the closed-world data used for training the embedding model.

For more details, please refer to our paper at https://arxiv.org/html/2305.12039v2!