Threshold-Consistent Margin Loss for Open-World Deep Metric Learning





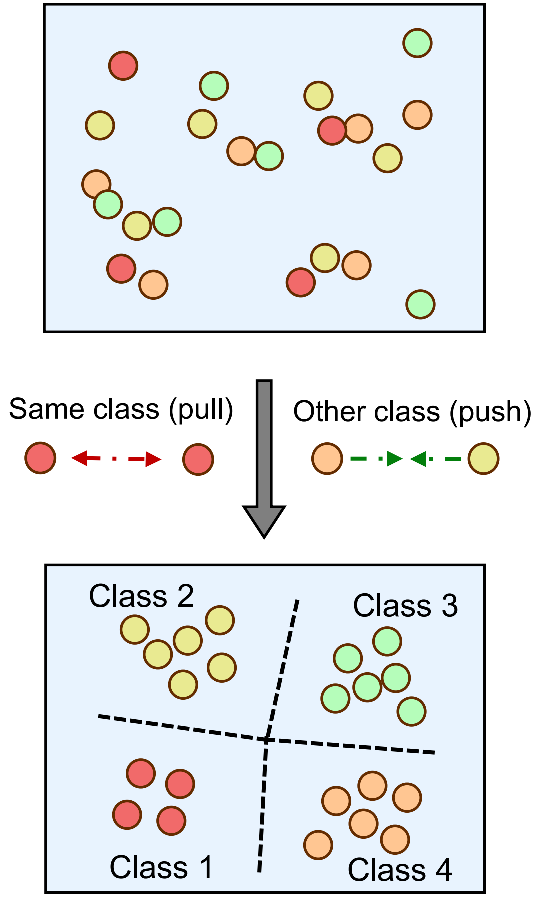


Qin Zhang Linghan Xu Qingming Tang Jun Fang Ying Nian Wu Joseph Tighe Yifan Xing

amazon Rekognition

Deep Metric Learning (DML)

Original data space



Learned metric space

Face recognition

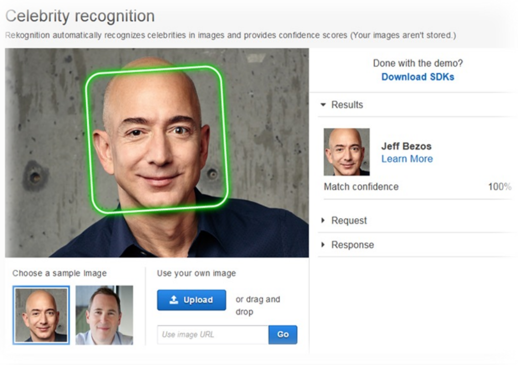


Image retrieval

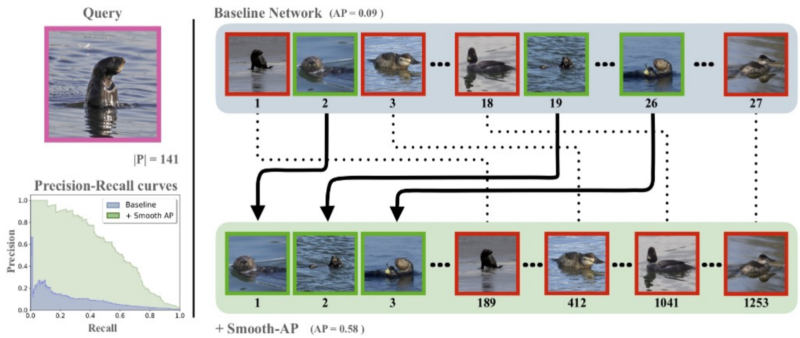
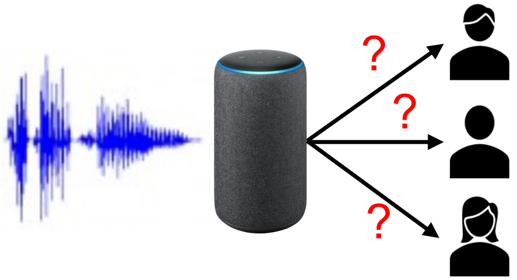


image source: https://arxiv.org/pdf/2007.12163.pdf

Speaker identification



Multimodal retrieval

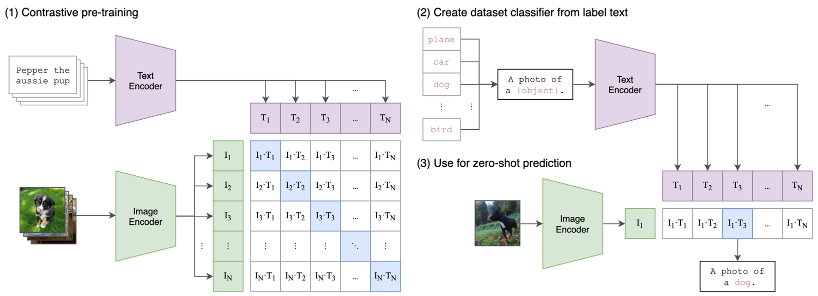
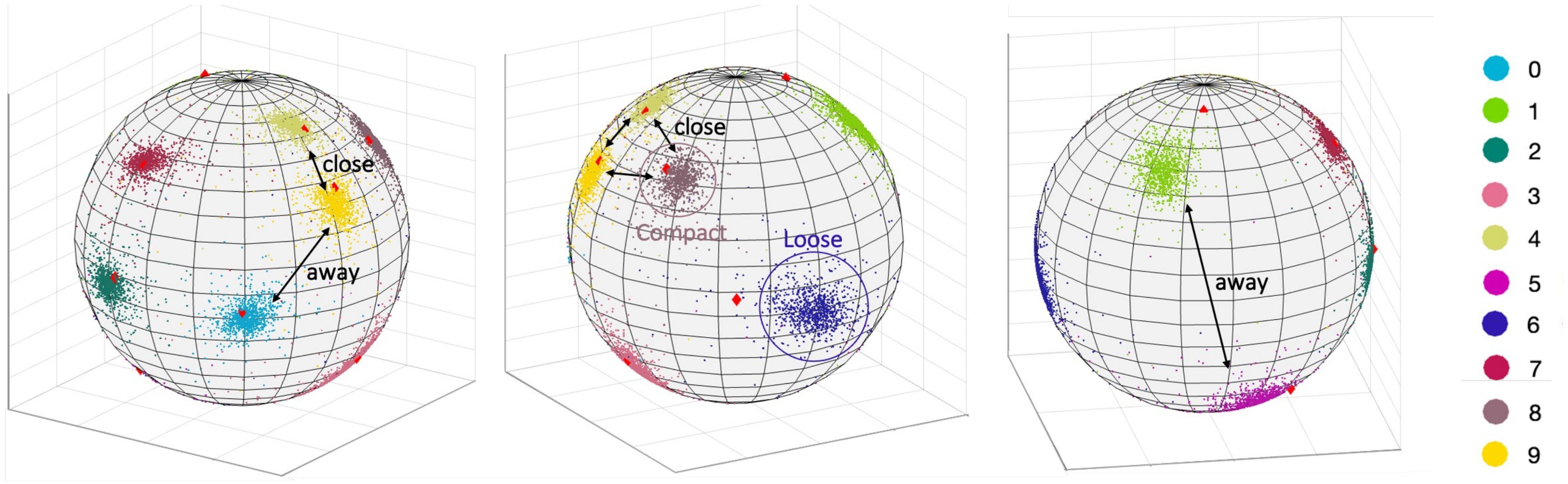


image source: https://arxiv.org/pdf/2103.00020.pdf

Broad Applications

DML Suffers from Inconsistencies in Metric Structures

- However, DML suffers from inconsistent metric structures across classes because standard DML losses do not explicitly ensure uniform intra-class compactness and consistent inter-class separation.
  - A toy example for MNIST handwritten digit dataset:



(a) Front view (b) Side view (c) Back view

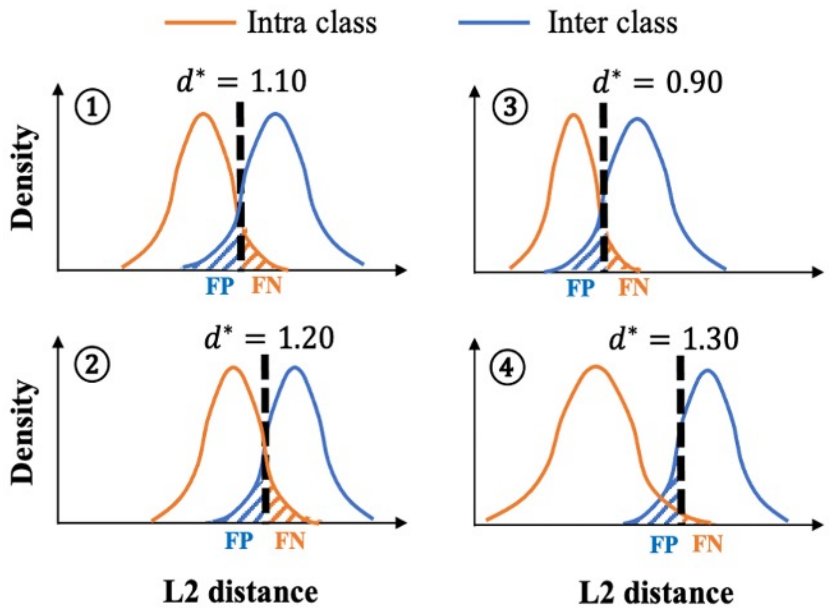
High DML Accuracy ≠ High User Experience

- Consider a scenario with diverse users representing different classes, each having different intra-class and inter-class metric structures. For retrieval/verification applications, these users would require distinct L2 distance thresholds to achieve targeted False Accept Rate or False Reject Rate accuracy metrics.

Diverse classes

Embedding model

Optimize for a target FPR



Inconsistent Thresholds!

1 $d^* = 1.10$
2 $d^* = 1.20$
3 $d^* = 0.90$
4 $d^* = 1.30$

Operating-Point-Inconsistency Score (OPIS)

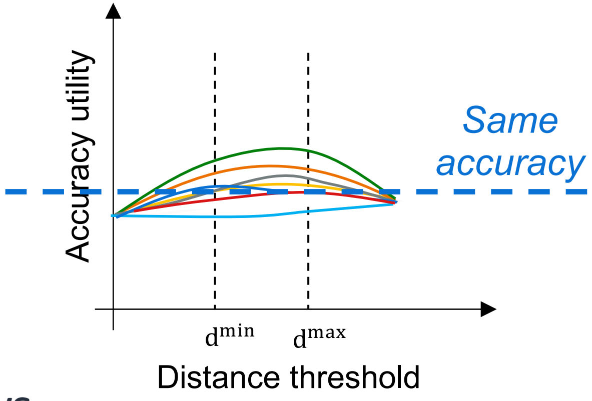
- We propose the OPIS metric: for a test set containing T members, OPIS quantifies the variance in accuracy utility (denoted as U) within a predefined distance threshold range $[\mathrm{d}^{min}, \mathrm{d}^{max}]$ across all members:

Utility of test member i

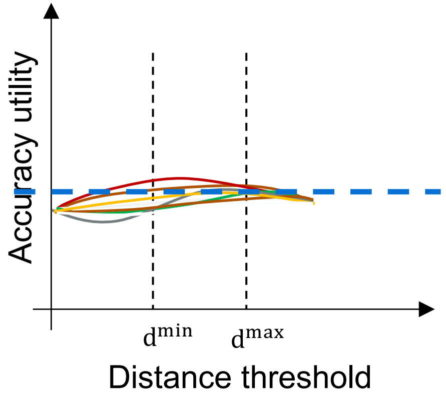
Average utility of all test members

$$\mathrm{OPIS} = \underbrace{\frac{1}{\mathrm{d}^{\max} - \mathrm{d}^{\min}}}_{\text{Normalize by range}} \times \underbrace{\frac{\sum_{i=1}^{T} \int_{\mathrm{d}^{\min}}^{\mathrm{d}^{\max}} \|\mathrm{U_i}(\mathrm{d}) - \overline{\mathrm{U}}(\mathrm{d})\|^2 \, d\mathrm{d}}{T}}_{\text{Variance}}$$

High OPIS



Low OPIS

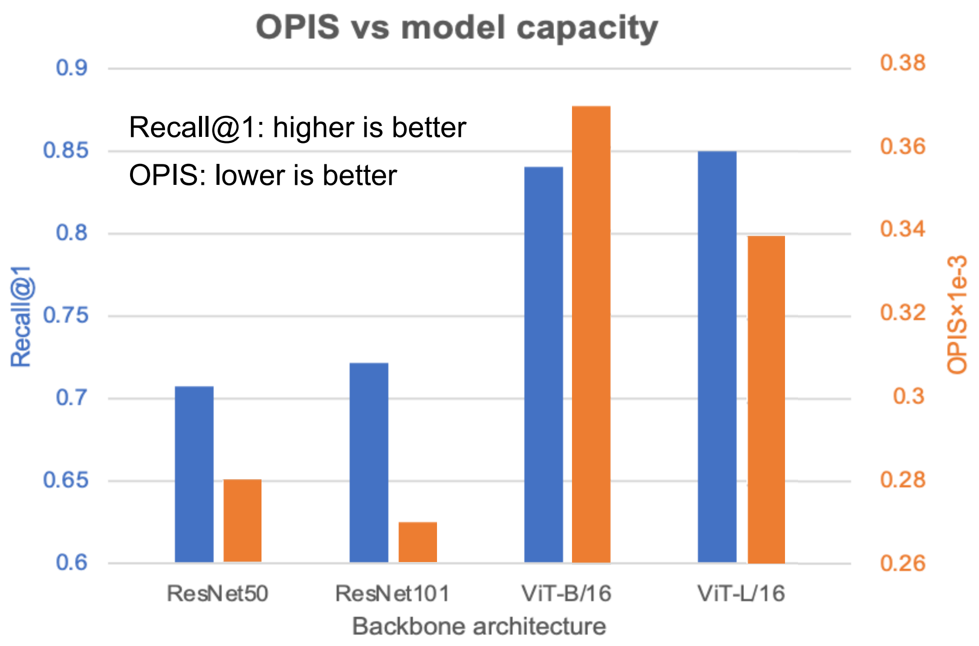


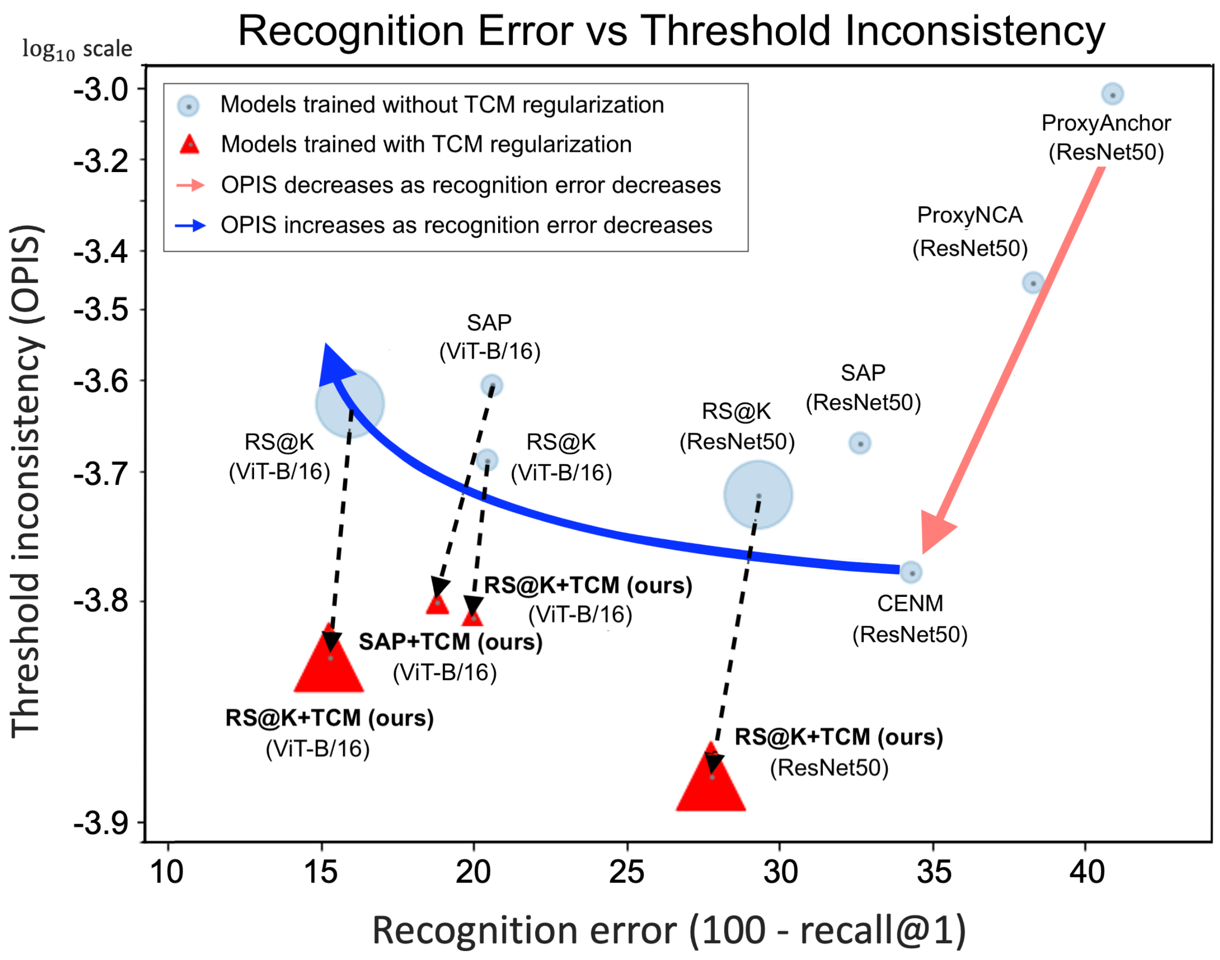
A few notes:

1. OPIS is a metric orthogonal to accuracy.
2. OPIS only makes sense when the are sufficient number of test members whose utility scores in the calibration range are statistically significant.
3. The test members can be different datasets or classes. For different classes, they should have similar concept granularities.

High DML Accuracy ≠ High Threshold Consistency

- Although it is commonly believed that larger DML models enhance intra-class compactness and inter-class separation, which benefits both model discriminability and threshold consistency, ***our observations yield mixed results***.



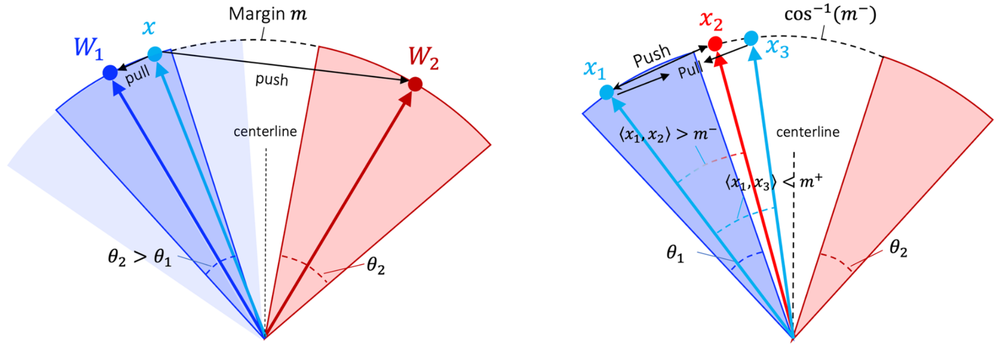


Threshold-Consistent Margin Loss (TCM)

- We propose the Threshold-Consistent Margin (TCM) Loss for improving threshold consistency in DML:

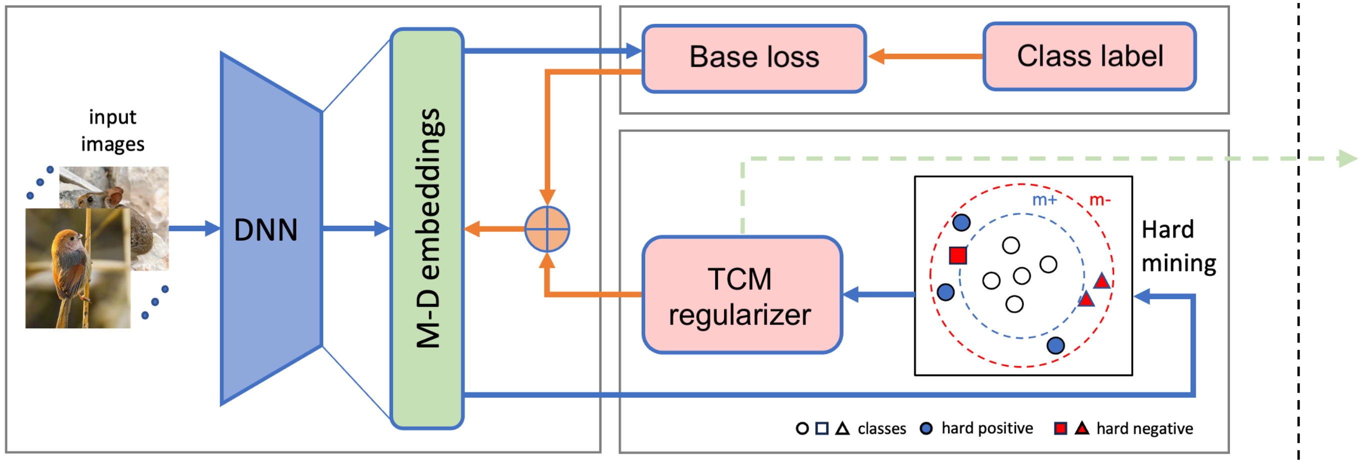
Positive margin Negative margin

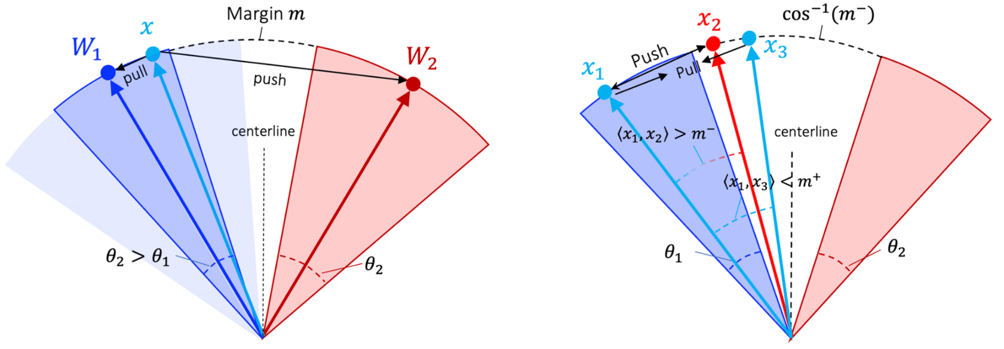
$$L_{\mathrm{TCM}} = \underbrace{\frac{\sum_{s \in S^+} (m^+ - s) \cdot 1_{s \le m^+}}{\sum_{s \in S^+} 1_{s \le m^+}}}_{\text{Penalty to hard positive pairs}} + \underbrace{\frac{\sum_{s \in S^-} (s - m^-) \cdot 1_{s \ge m^-}}{\sum_{s \in S^-} 1_{s \ge m^-}}}_{\text{Penalty to hard negative pairs}}$$



(i) Margin-based Softmax (ii) Pairwise TCM

- TCM can be combined with any base DML loss:





(i) Margin-based Softmax (ii) Pairwise TCM

aws

Results: TCM Significantly Improves Threshold Consistency

- When used as a regularization loss alongside SoTA losses such as Smooth-AP loss and Recall@K surrogate loss, TCM not only achieves competitive accuracy improvements (+1.05% in Recall@1) but also significantly enhances threshold consistency, achieving a relative reduction of up to 77.3%.

Sample images	Benchmark	Arch^{dim}	$L_{\text{base}} + L_{\text{TCM}}$	BS	OPIS $\times 10^{-3}\downarrow$	10%-OPIS $\times 10^{-3}\downarrow$	R@1 $\uparrow$	Previous SOTA with ImageNet pretraining
	iNaturalist-2018	ResNet50^{512}	SAP + TCM	384	0.17 ↓0.16 (−48.5%)	1.77 ↓2.83 (−61.5%)	69.1 ↑1.7	R@1: 83.9 ViT-B/16 (Patel et al., 2022)
			RS@k + TCM	4000	0.11 ↓0.17 (−60.7%)	1.25 ↓2.49 (−66.6%)	72.2 ↑1.5	
		ViT-B/16^{512}	SAP + TCM	384	0.20 ↓0.19 (−48.7%)	2.81 ↓2.40 (−46.1%)	81.2 ↑1.8	
			RS@k + TCM	4000	0.17 ↓0.20 (−54.1%)	2.03 ↓5.63 (−73.5%)	**84.8** ↑0.9	
	Stanford Online Product	ResNet50^{512}	SAP + TCM	384	0.06 ↓0.11 (−64.7%)	0.52 ↓1.17 (−69.2%)	82.7 ↑2.9	R@1: 88.0 ViT-B/16 (Patel et al., 2022)
			RS@k + TCM	4000	0.07 ↓0.03 (−30.1%)	0.74 ↓0.12 (−14.0%)	83.3 ↑0.6	
		ViT-B/16^{512}	SAP + TCM	384	0.04 ↓0.01 (−25.4%)	0.33 ↓0.11 (−25.0%)	87.3 ↑0.2	
			RS@k + HMC	4000	0.04 ↓0.00 (−3.7%)	0.38 ↓0.08 (−17.4%)	**88.4** ↑0.4	
	CUB-200-2011	ResNet50^{512}	SAP + TCM	384	0.11 ↓0.04 (−26.7%)	1.00 ↓0.43 (−30.1%)	80.8 ↑1.0	R@1: 85.7 ViT-S/16 (Kim et al., 2023)
			RS@k + TCM	384	0.10 ↓0.12 (−54.5%)	0.91 ↓1.04 (−53.3%)	80.0 ↑0.7	
		ViT-B/16^{512}	SAP + TCM	384	0.07 ↓0.14 (−66.7%)	0.58 ↓1.08 (−65.1%)	**88.4** ↑0.0	
			RS@k + TCM	384	0.10 ↓0.34 (−77.3%)	0.91 ↓2.66 (−74.5%)	87.6 ↓0.1	
	Cars-196	ResNet50^{512}	SAP + TCM	384	0.39 ↓0.06 (−13.3%)	3.33 ↓1.24 (−27.1%)	89.6 ↑2.7	R@1: **91.3** DINO (Kim et al., 2023)
			RS@k + TCM	392	0.45 ↓0.02 (−4.3%)	2.93 ↓0.65 (−18.2%)	89.7 ↓0.2	
		ViT-B/16^{512}	SAP + TCM	384	0.54 ↓0.66 (−55.2%)	0.83 ↓1.79 (−68.3%)	87.8 ↑0.7	
			RS@k + TCM	392	0.60 ↓0.37 (−38.1%)	0.98 ↓1.73 (−63.8%)	87.7 ↑0.8	

The 10%-OPIS metric quantifies the utility performance disparity between the best-performing 10% of classes and the worst-performing 10%.

Takeaways for "Threshold-Consistent Margin Loss for Open-World Deep Metric Learning"

- We find that achieving high accuracy levels in a DML model does not automatically guarantee threshold consistency, which may directly impact user experience in real-world deployment environment.
- To quantify the severity of this problem, we propose a novel variance-based metric called Operating-Point-Inconsistency-Score (OPIS) that quantifies the variance in the operating characteristics across classes.
- Using the OPIS metric, we observe a Pareto frontier in the high-accuracy regime, where existing DML methods to improve accuracy often lead to degradation in threshold consistency.
- To address this trade-off, we introduce the Threshold-Consistent Margin (TCM) loss, a simple yet effective regularization technique that promotes uniformity in representation structures across classes by selectively penalizing hard sample pairs.
- Extensive experiments demonstrate TCM's effectiveness in enhancing threshold consistency while preserving accuracy, simplifying the threshold selection process in practical DML settings.

For more details, please refer to our paper at https://openreview.net/pdf?id=vE5MyzpP92!
The corresponding code will be released soon.