JOHNS HOPKINS UNIVERSITY

arXiv

Enhancing Adversarial Robustness for Deep Metric Learning

Mo Zhou [1] Vishal M. Patel [1]

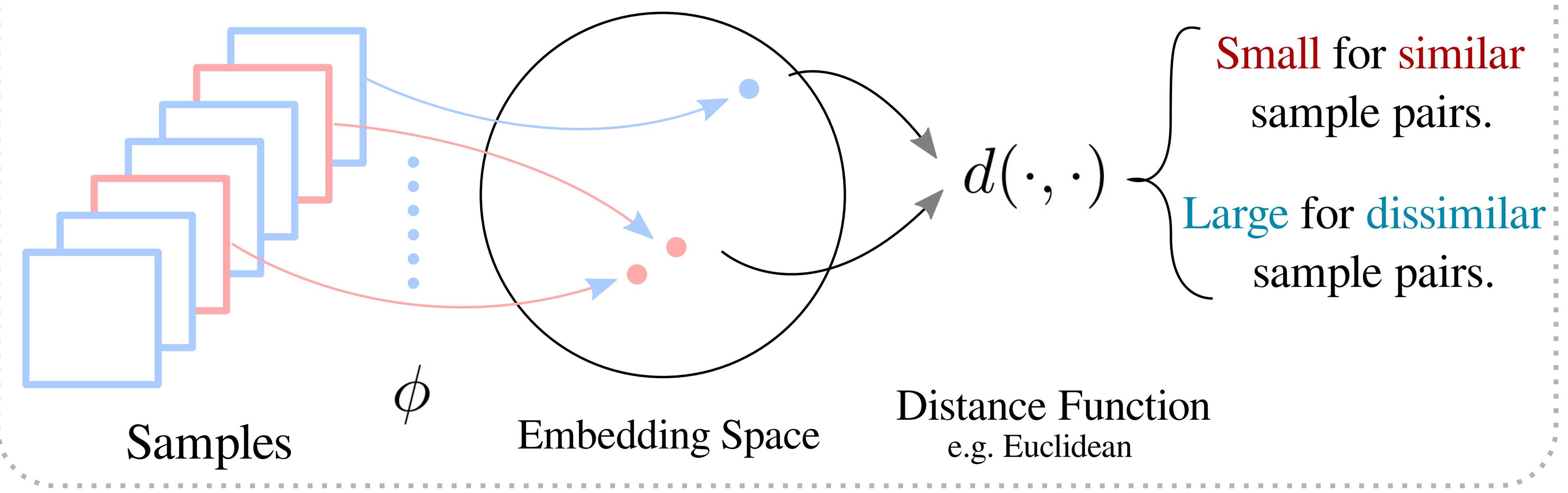
[1] Johns Hopkins University

Github

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Deep Metric Learning

Goal: Learn a mapping function $\phi: \mathcal{X} \mapsto \Phi \subseteq \mathbb{R}^D$



Insights / Motivations

- Deep metric learning (DML) model is vulnerable to adversarial attacks, which results in unexpected retrieval results.
- Leads to safety and security concerns to DML applications.

Embedding-Shifted Triplet (EST) [9] adopts adversarial counterparts of $\boldsymbol{a}, \boldsymbol{p}, \boldsymbol{n}$ with maximum embedding move distance off their original locations, *i.e.*,

$$L_{\text{EST}} = L_{\text{T}}(\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{p}}, \tilde{\boldsymbol{n}}; \gamma) \tag{1}$$

where $\tilde{\boldsymbol{a}} = \phi(\boldsymbol{A}+\boldsymbol{r}^*)$, and $\boldsymbol{r}^* = \arg\max_{\boldsymbol{r}} d_\phi(\boldsymbol{A}+\boldsymbol{r}, \boldsymbol{A})$. The $\tilde{\boldsymbol{p}}$ and $\tilde{\boldsymbol{n}}$ are obtained similarly.

Anti-Collapse Triplet (ACT) [10] "collapses" the embedding vectors of positive and negative sample, and enforces the model to separate them apart, *i.e.*,

$$L_{\text{ACT}} = L_{\text{T}}(\boldsymbol{a}, \overrightarrow{\boldsymbol{p}}, \overleftarrow{\boldsymbol{n}}; \gamma), \tag{2}$$
$$[\overrightarrow{\boldsymbol{p}}, \overleftarrow{\boldsymbol{n}}] = [\phi(\boldsymbol{P}+\boldsymbol{r}_p^*), \phi(\boldsymbol{N}+\boldsymbol{r}_n^*)] \tag{3}$$
$$[\boldsymbol{r}_p^*, \boldsymbol{r}_n^*] = \arg\min_{\boldsymbol{r}_p, \boldsymbol{r}_n} d_\phi(\boldsymbol{P}+\boldsymbol{r}_p, \boldsymbol{N}+\boldsymbol{r}_n). \tag{4}$$

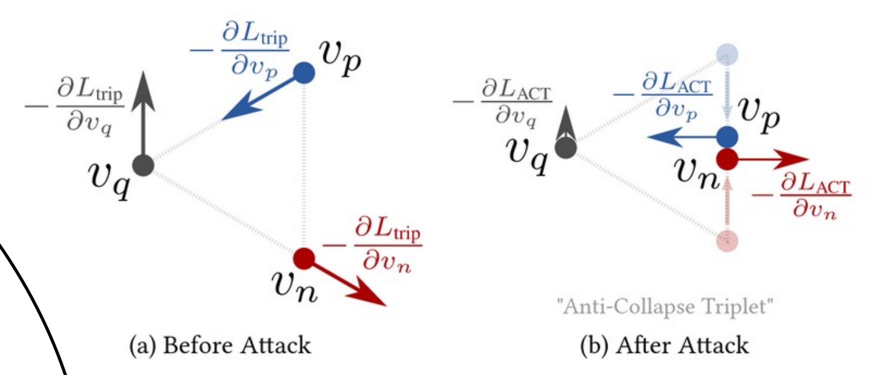
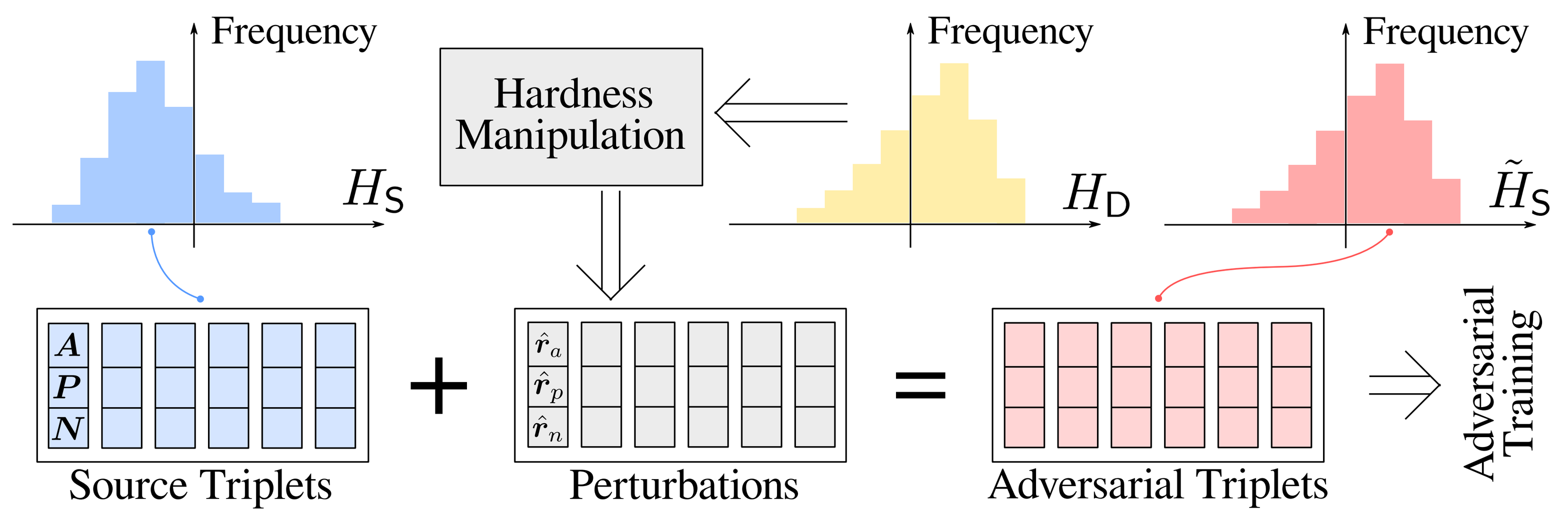


Fig. 4. Gradient Direction of Anti-Collapse Triplet (ACT). ACT does not suffer from misleading gradient or inefficient mini-batch exploitation. It does not create excessively hard adversarial example neither.

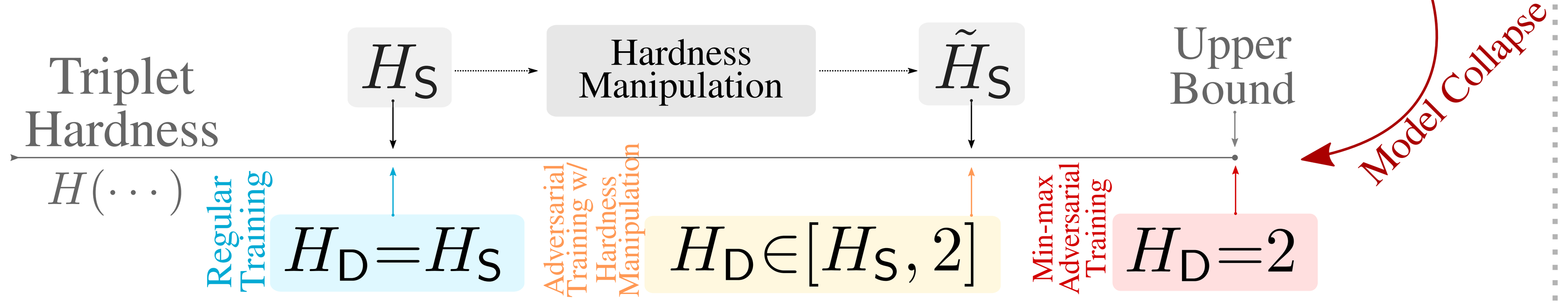
- Existing defense methods (i.e., EST and ACT) learn from relatively weak adversaries in order to avoid model collapse.
- And suffer from low efficiency in adversarial training, low performance on benign examples, and insufficient robustness.

Contribution 1

Hardness Manipulation



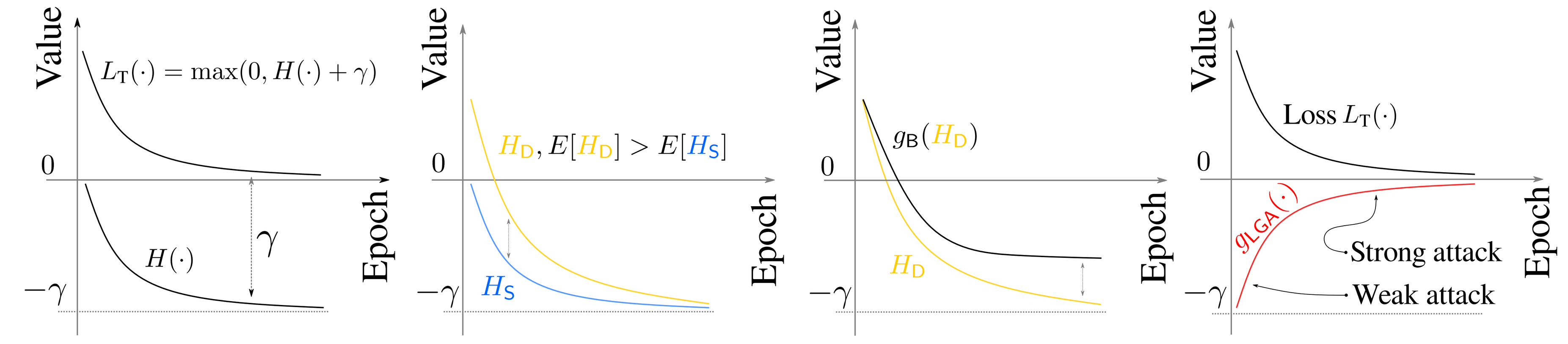
- Perturb a given image triplet to a specific hardness level, e.g., from semi-hard to soft-hard, instead of as hard as possible.



- Can be seen as a flexible and dynamic "interpolation" between regular training and Madry min-max adversarial training.
- Efficiently create adversarial examples for training and do not easily lead to model collapse.

Contribution 2

Linear Gradual Adversary

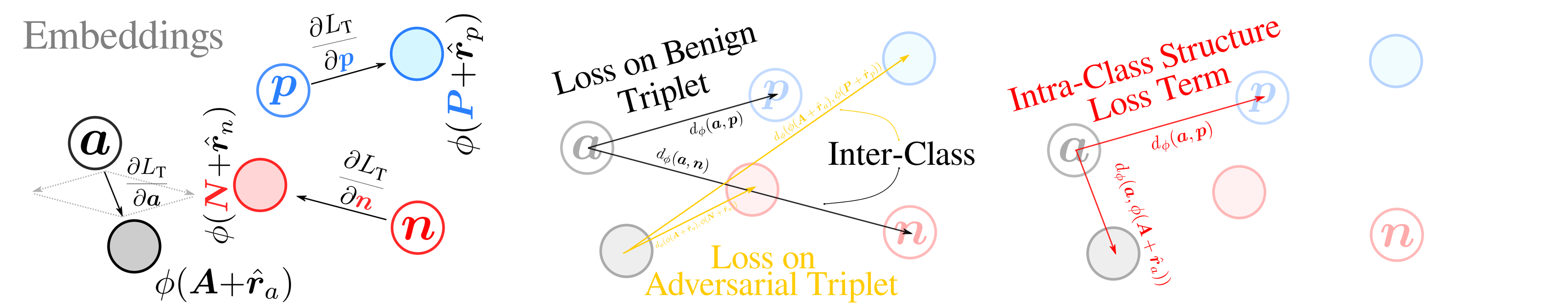


(a) Relationship between loss and hardness when $L_\text{T}(\cdot){>}0$ (b) Shrinked $E[H_\mathsf{D} - H_\mathsf{S}]$ value leads to weak attack (c) Slightly Boosting H_D for slightly stronger attack (d) Linear Gradual Adversary for balanced training

- Gradually and dynamically increases hardness level to balance the model robustness and performance on benign examples.

Contribution 3

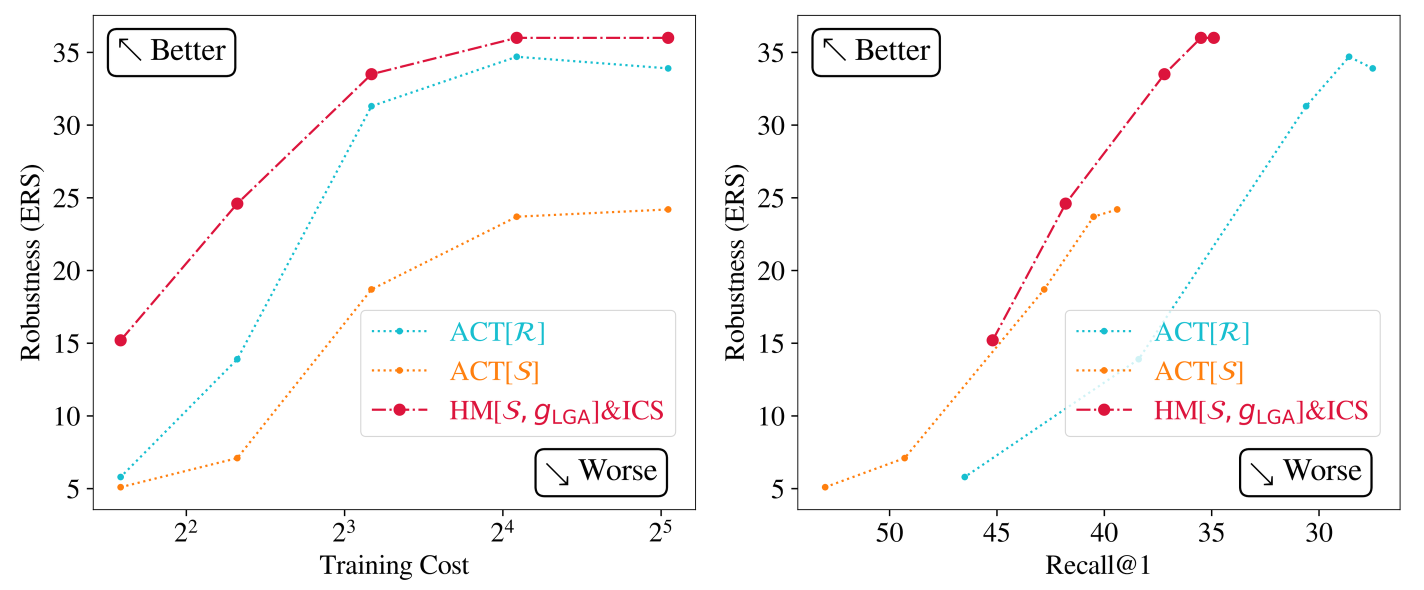
Intra-Class Structure Loss



(a) Benign & Adversarial Triplets = Sextuplet (b) Loss on Benign and Adversarial Triplets (c) Intra-Class Structure Loss Term L_ICS

- Enforcing intra-class structure is neglected by existing defenses.
- Some attacks happen within a class.

Results



- Overwhelmingly outperform SoTA in efficiency of adversarial training, performance on benign examples, and robustness.
- Comparison on:
  (1) CUB-200-2011
  (2) Cars-196
  (2) Stanford Online Product

This table is vector graphics cropped from my PDF instead of bitmap screenshot.
Zoom in and see the numbers.

Dataset	Defense	η	Benign Example R@1↑	R@2↑	mAP↑	NMI↑	White-Box Attacks for Robustness Evaluation CA+↑	CA-↓	QA+↑	QA-↓	TMA↓	ES:D↓	ES:R↑	LTM↑	GTM↑	GTT↑	ERS↑
CUB	N/A[$\mathcal{R}$]	N/A	53.9	66.4	26.1	59.5	0.0	100.0	0.0	99.9	0.883	1.762	0.0	0.0	14.1	0.0	3.8
	EST[$\mathcal{R}$] [53]	8	37.1	47.3	20.0	46.4	0.5	97.3	0.5	91.3	0.875	1.325	3.9	0.4	14.9	0.0	7.9
	ACT[$\mathcal{R}$] [55]	8	30.6	40.1	16.5	45.6	13.7	46.8	12.6	39.3	0.547	0.902	13.6	9.8	21.9	1.3	31.3
	HM[$\mathcal{S}$, g_LGA]	8	38.0	48.3	21.8	49.3	12.7	46.4	11.6	39.9	0.567	0.783	16.8	11.9	27.9	1.4	32.4
	HM[$\mathcal{S}$, g_LGA]&ICS	8	37.2	47.8	21.4	48.4	12.9	40.9	14.7	33.7	0.806	0.487	17.1	13.2	26.3	2.3	33.5
	EST[$\mathcal{R}$] [53]	32	8.5	13.0	2.6	25.2	2.7	97.9	0.4	97.3	0.848	1.576	1.4	0.0	4.0	0.0	5.3
	ACT[$\mathcal{R}$] [55]	32	27.5	38.2	12.2	43.0	15.5	37.7	15.1	32.2	0.472	0.821	11.1	9.4	14.9	1.0	33.9
	HM[$\mathcal{S}$, g_LGA]	32	36.5	46.7	21.0	48.6	14.7	39.6	15.6	34.2	0.523	0.736	16.5	15.0	26.7	2.9	35.9
	HM[$\mathcal{S}$, g_LGA]&ICS	32	34.9	45.0	19.8	47.1	15.5	37.7	16.6	30.9	0.753	0.506	17.9	16.7	27.3	2.9	36.0
CARS	N/A[$\mathcal{R}$]	N/A	62.5	74.0	23.8	57.0	0.2	100.0	0.1	99.6	0.874	1.816	0.0	0.0	13.4	0.0	3.6
	EST[$\mathcal{R}$] [53]	8	57.1	68.4	30.3	47.7	0.1	99.9	0.1	98.1	0.902	1.681	0.7	0.2	15.4	0.0	4.4
	ACT[$\mathcal{R}$] [55]	8	46.8	58.0	23.4	45.5	19.3	33.1	20.3	32.3	0.413	0.760	18.4	15.0	28.6	1.2	39.8
	HM[$\mathcal{S}$, g_LGA]	8	63.2	73.7	36.8	53.5	15.3	32.0	17.9	33.9	0.463	0.653	23.4	28.5	44.6	5.8	42.4
	HM[$\mathcal{S}$, g_LGA]&ICS	8	61.7	72.6	35.5	51.8	21.0	23.3	23.1	22.2	0.698	0.415	31.2	38.0	47.8	9.6	47.9
	EST[$\mathcal{R}$] [53]	32	30.7	41.0	5.6	31.8	1.2	98.1	0.4	91.8	0.880	1.281	2.9	0.7	8.2	0.0	7.3
	ACT[$\mathcal{R}$] [55]	32	43.4	54.6	11.8	42.9	18.0	32.3	17.5	30.5	0.383	0.763	16.3	15.3	20.7	1.6	38.6
	HM[$\mathcal{S}$, g_LGA]	32	62.3	72.5	35.3	52.7	17.4	28.2	18.2	28.8	0.426	0.613	27.1	30.7	42.3	7.9	44.9
	HM[$\mathcal{S}$, g_LGA]&ICS	32	60.2	71.6	33.9	51.2	19.3	25.9	19.6	25.7	0.650	0.446	30.3	36.7	46.0	8.8	46.0
SOP	N/A[$\mathcal{R}$]	N/A	62.9	68.5	39.2	87.4	0.1	99.3	0.2	99.1	0.845	1.685	0.0	0.0	6.3	0.0	4.0
	EST[$\mathcal{R}$] [53]	8	52.7	58.5	30.1	85.7	6.4	69.7	3.9	64.6	0.611	1.053	3.8	2.2	10.2	1.3	19.0
	ACT[$\mathcal{R}$] [55]	8	45.3	50.6	24.1	84.7	24.8	10.7	25.4	8.2	0.321	0.485	15.4	17.7	25.1	11.3	49.5
	HM[$\mathcal{S}$, g_LGA]	8	49.0	54.1	26.4	85.0	29.9	4.7	31.6	3.6	0.455	0.283	39.3	40.9	38.8	43.0	61.7
	HM[$\mathcal{S}$, g_LGA]&ICS	8	48.3	53.4	25.7	84.9	32.5	4.8	32.4	3.5	0.586	0.239	38.6	39.8	38.3	44.5	61.2
	EST[$\mathcal{R}$] [53]	32	46.0	51.4	24.5	84.7	12.5	43.6	10.6	34.8	0.468	0.830	9.6	7.2	17.3	3.8	31.7
	ACT[$\mathcal{R}$] [55]	32	47.5	52.6	25.5	84.9	24.1	10.5	22.7	9.4	0.253	0.532	21.2	21.6	27.8	15.3	50.8
	HM[$\mathcal{S}$, g_LGA]	32	47.7	52.7	25.3	84.8	30.6	4.7	31.2	3.5	0.466	0.266	38.6	40.3	38.6	44.3	61.8
	HM[$\mathcal{S}$, g_LGA]&ICS	32	46.8	51.7	24.5	84.7	32.0	4.2	33.7	3.0	0.606	0.207	39.1	39.8	37.9	45.6	61.6

Table 6. Comparison of our defense with the state-of-the-art methods on commonly used DML datasets.

η: number of PGD iterations for adversarial example, lower=better
R@1: recall @ 1 metric for deep metric learning, higher=better
ERS: empirical robustness score, higher=better