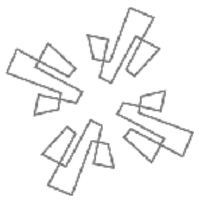
ACM Multimedia Asia 2022

# INFORMATIVE SAMPLE-AWARE PROXY FOR DEEP METRIC LEARNING

*Aoyu Li, Ikuro Sato, Kohta Ishikawa, Rei Kawakami, Rio Yokota*

*Presenter: Aoyu Li*

*GSIC, Tokyo Institute of Technology*



DENSO IT LAB

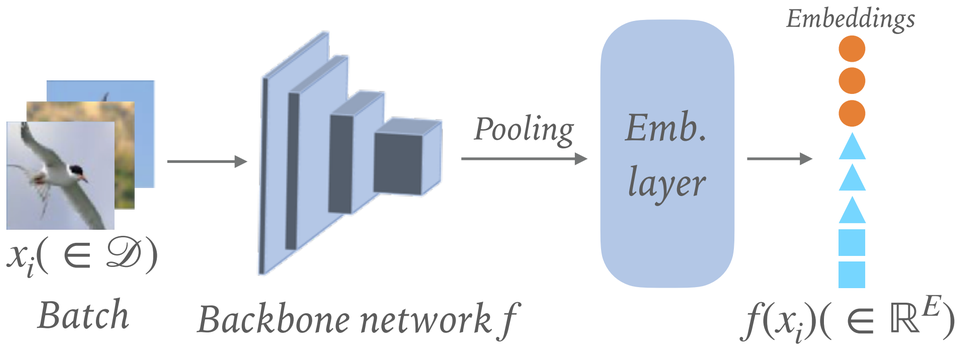
# Introduction to Deep (Distance) Metric Learning

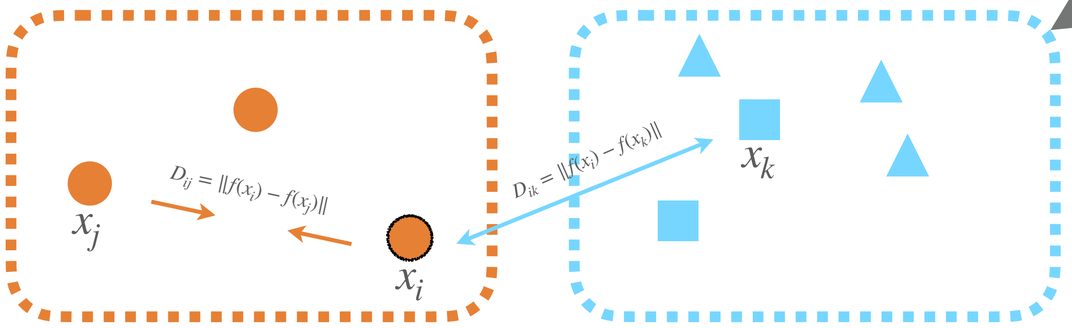
- Deep Metric Learning (DML): supervised representation learning such that

Data from the same class: → Pull together

Data from different class: → Push apart

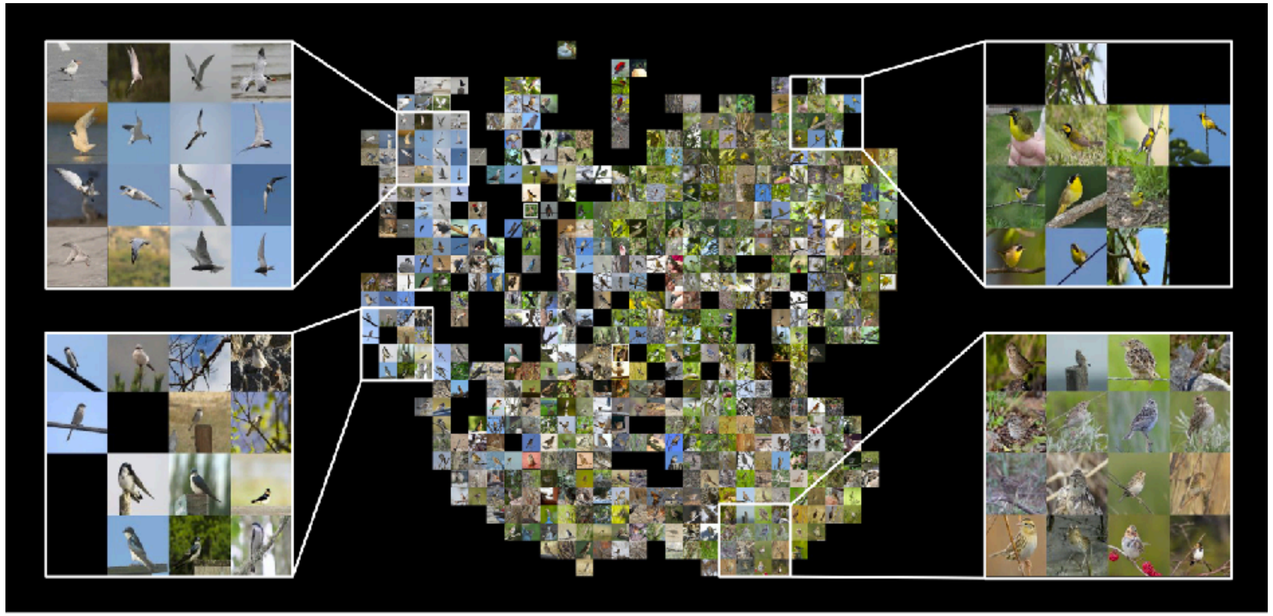
*Euclidean distance or cosine similarity*




*Positives:* $y_j = y_i$

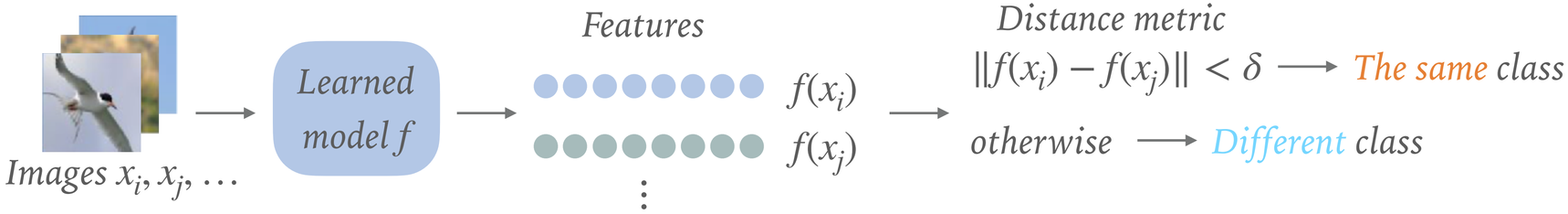
*Negatives:* $y_k \neq y_i$



*Embedding space (visualized with t-SNE [24])*

# Why Metric Learning ?

- Tasks cannot be treated as classification when:
  - The number of classes is indefinite.
  - Images belonging to each class cannot be obtained in advance.
  - With DML:

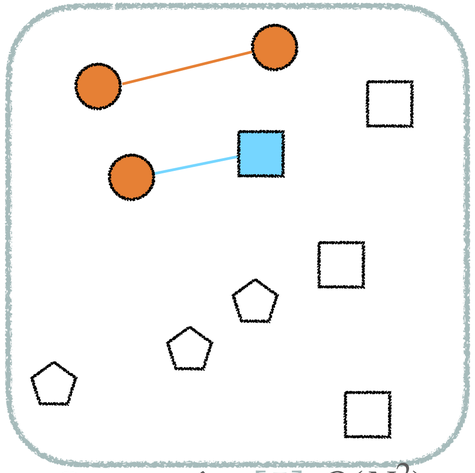



- Application:

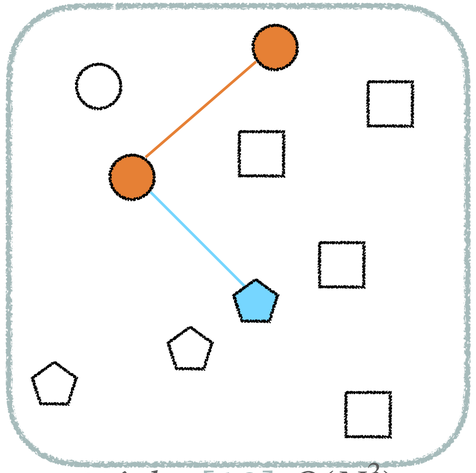
  Image Retrieval, Face Recognition, Few-shot Learning, etc.

# Types of DML Methods

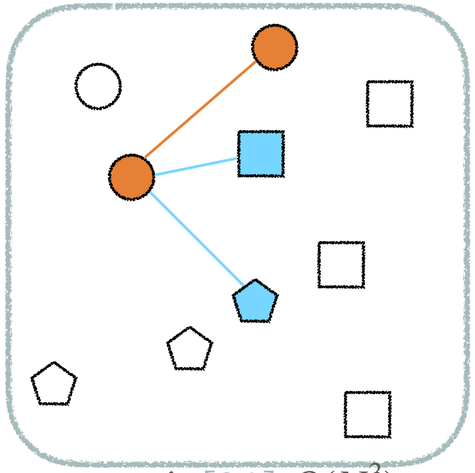
- Pair-based loss



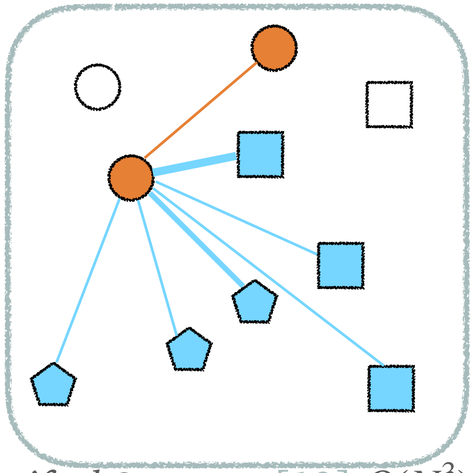
*Contrastive* [5] $O(N^2)$



*Triplet* [18] $O(N^3)$



*N-Pair* [21] $O(N^3)$

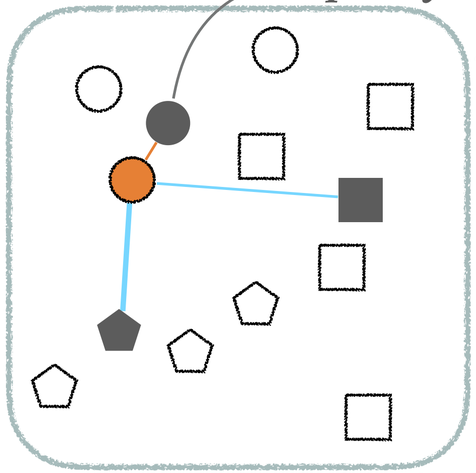


*Lifted Structure* [12] $O(N^3)$

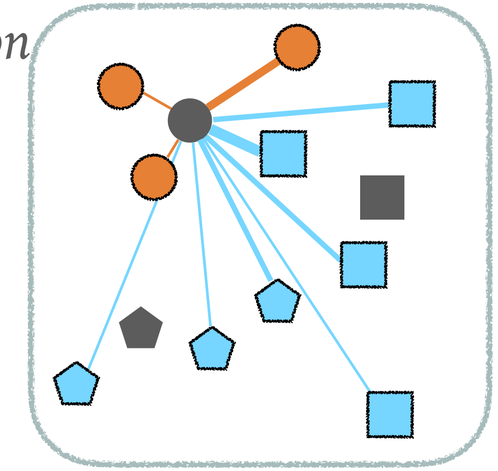
- Proxy-based loss

*proxy: centroid of a class*

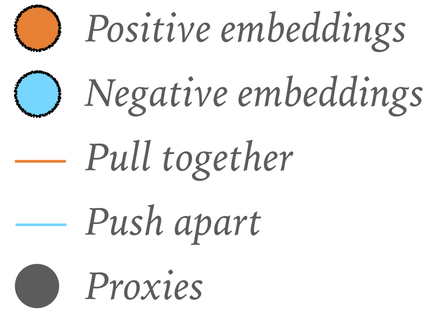
*source of global information*



*Proxy-NCA* [10] $O(NC)$



*Proxy-Anchor* [9] $O(NC)$



# General Pair Weighting (X. Wang+, 2019 [11]) for DML Loss

A DML loss can be formulated as $\mathscr{L} = \sum_i \sum_{j \neq i} L(S_{ij}, y_i, y_j)$.

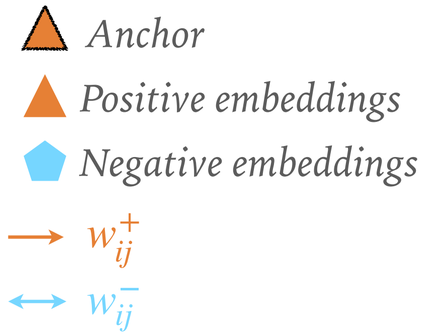
Let $x_i \in \mathbb{R}^D$ be a row data sample, label $y_i \in \{1,2,\ldots,C\}$;

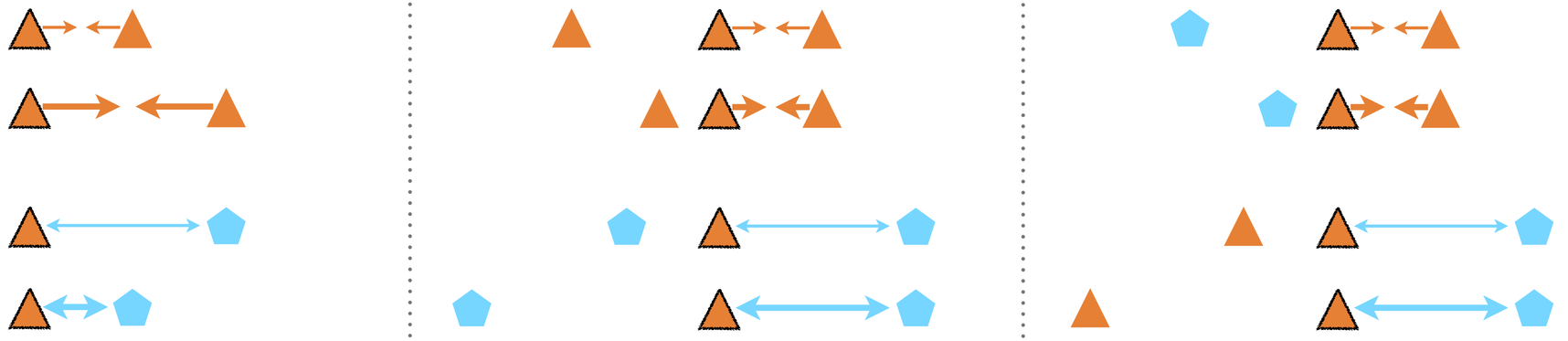
Encoder $f(\cdot\,;\theta) : \mathbb{R}^D \to \mathbb{R}^E$; Similarity of two samples: $S_{ij} = < f(x_i;\theta), f(x_j;\theta) >$ ;

*negative pairs* *positive pairs*

$$\frac{\partial \mathscr{L}}{\partial \theta} = \sum_i \left( \sum_{j:y_j \neq y_i} \frac{\partial L}{\partial S_{ij}} \frac{\partial S_{ij}}{\partial \theta} + \sum_{j:y_j = y_i} \frac{\partial L}{\partial S_{ij}} \frac{\partial S_{ij}}{\partial \theta} \right) := \sum_i \left( \sum_{j:y_j \neq y_i} w_{ij}^{-} \frac{\partial S_{ij}}{\partial \theta} - \sum_{j:y_j = y_i} w_{ij}^{+} \frac{\partial S_{ij}}{\partial \theta} \right)$$

# Gradient Weights & Types of Hardness



*Self hardness* | *Relative hardness* | *Class hardness*

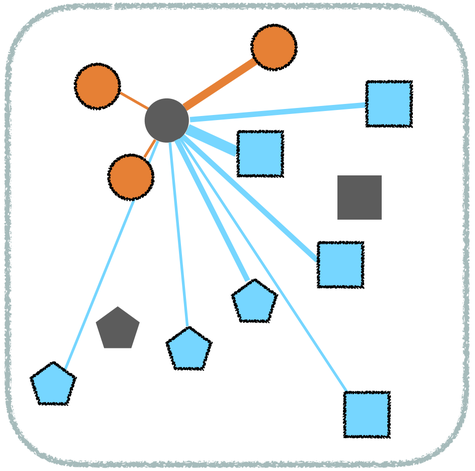
| | Self hardness | Relative hardness | Class hardness |
|---|---|---|---|
| Contrastive [5] | ✓ | ✗ | ✗ |
| Triplet [18] | ✗ | ✗ | ✓ |
| Lifted Structure [12] | ✗ | ✓ | ✗ |
| MS [11] | ✓ | ✓ | ✗ |
| Norm. SoftMax [23] | ✗ | ✗ | ✓ |
| Proxy-Anchor [9] | ✓ | ✓ | ✗ |

# Gradient Weights & Types of Hardness

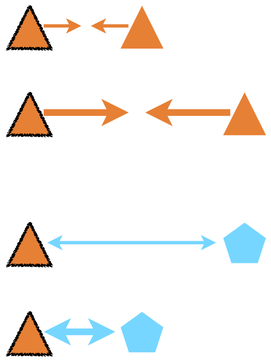
Proxy

Positive embeddings

Negative embeddings



Proxy-Anchor [9] $O(NC)$

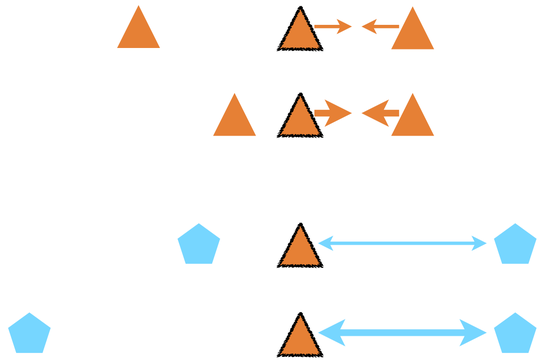


## Self hardness

e.g. Binomial Deviance [13]

$$w_{ij}^{+} = \frac{1}{P_i} \frac{\alpha \exp\{\alpha(\lambda - S_{ij})\}}{1 + \exp\{\alpha(\lambda - S_{ij})\}}$$

$$w_{ij}^{-} = \frac{1}{N_i} \frac{\beta \exp\{\beta(S_{ij} - \lambda)\}}{1 + \exp\{\beta(S_{ij} - \lambda)\}}$$

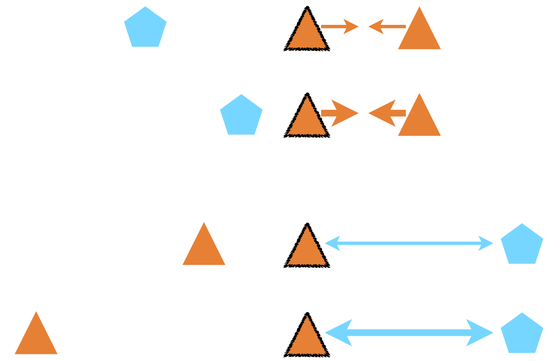


## Relative hardness

e.g. Lifted Structure [12]

$$w_{ij}^{+} = \frac{1}{\sum_{k:y_k=y_i} \exp\{S_{ij} - S_{ik}\}}$$

$$w_{ij}^{-} = \frac{1}{\sum_{k:y_k \neq y_i} \exp\{S_{ik} - S_{ij}\}}$$



## Class hardness

e.g. Normalized SoftMax [23]

$$w_{iy_i} = \alpha \left(1 - \frac{\exp\{\alpha S_{iy_i}\}}{\sum_c \exp\{\alpha S_{ic}\}}\right)$$

$$w_{ic}^{-} = \frac{\alpha \exp\{\alpha S_{ic}\}}{\sum_{c'} \exp\{\alpha S_{ic'}\}}$$

| | ✓ | ✓ | ✗ |
|---|---|---|---|
| Proxy-Anchor [9] | $L_{PA} = \frac{1}{\lvert P^+ \rvert} \sum_{p_c \in P^+} \log\left(1 + \sum_{i:y_i=c} \exp\{\alpha(\delta - S_{ic})\}\right) + \frac{1}{C} \sum_{c=1}^{C} \log\left(1 + \sum_{i:y_i \neq c} \exp\{\alpha(S_{ic} + \delta)\}\right)$ | $w_{ic}^{+} = \frac{1}{\lvert P^+ \rvert} \frac{\alpha \exp\{\alpha(\delta - S_{ic})\}}{1 + \sum_{j:y_j=c} \exp\{\alpha(\delta - S_{jc})\}}$ $w_{ic}^{-} = \frac{1}{C} \frac{\alpha \exp\{\alpha(S_{ic} + \delta)\}}{1 + \sum_{j:y_j \neq c} \exp\{\alpha(S_{jc} + \delta)\}}$ | |

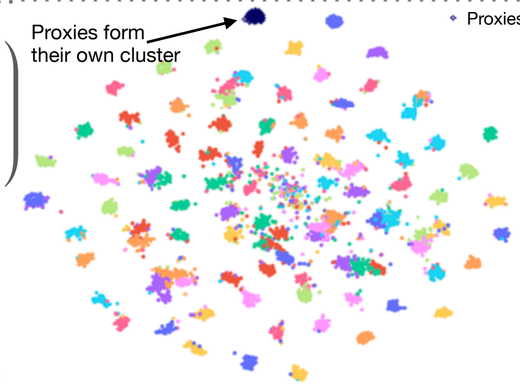
# Negative Impact of Pair Weighting Without the Class Hardness

- $L_{PA} = \frac{1}{|P^+|} \sum_{p_c \in P^+} \log\left(1 + \sum_{i:y_i=c} \exp\{\alpha(\delta - S_{ic})\}\right) + \frac{1}{C} \sum_{c=1}^{C} \log\left(1 + \sum_{i:y_i \neq c} \exp\{\alpha(S_{ic} + \delta)\}\right)$
- According to the K.K.T. condition, for the positive terms,

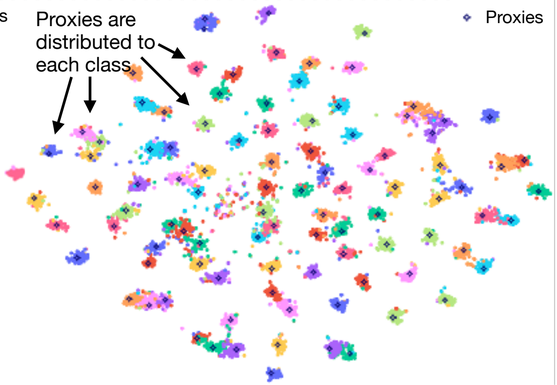
$$\log\left(1 + \sum_{i:y_i=c} e^{\alpha(\delta - S_{ic})}\right) = \max_{\mathscr{P}_c^+} \alpha \sum_{i:y_i=c} \mathscr{P}_c^+(i)(\delta - S_{ic}) + H(\mathscr{P}_c^+),$$

and the solution is $\mathscr{P}_c^+(i) = \frac{e^{\alpha(\delta - S_{ic})}}{1 + \sum_{j:y_j=c} e^{\alpha(\delta - S_{jc})}}$.

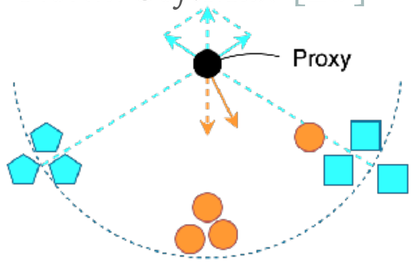
- Only optimizing two data distribution (positives, negatives) related to each proxy.
  - $w_{ic} = \frac{\partial L(S)}{\partial S_{ic}}$ works for all data points but also the proxy $p_c$ itself.
  - When **outlier** exists, the distribution would be sub-optimal.
  - Proxies form a cluster due to **strong repulsive forces** in the early learning stage.



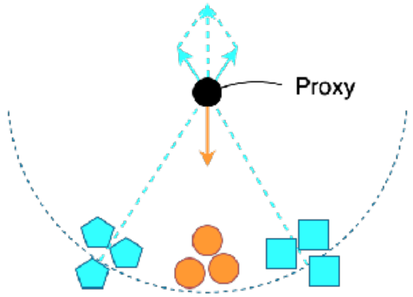

*Proxy-Anchor* [9]




*Norm. SoftMax* [23]



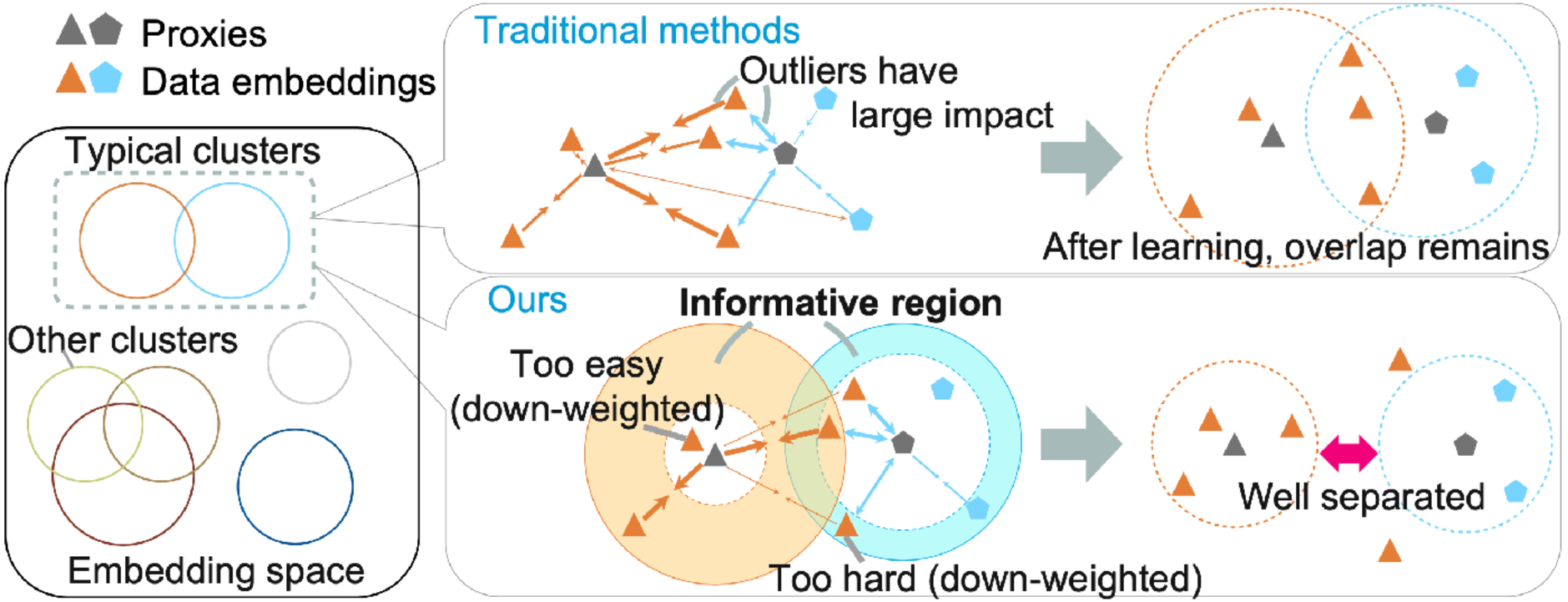

*Terms contain outlier*




*Compact at early stage*

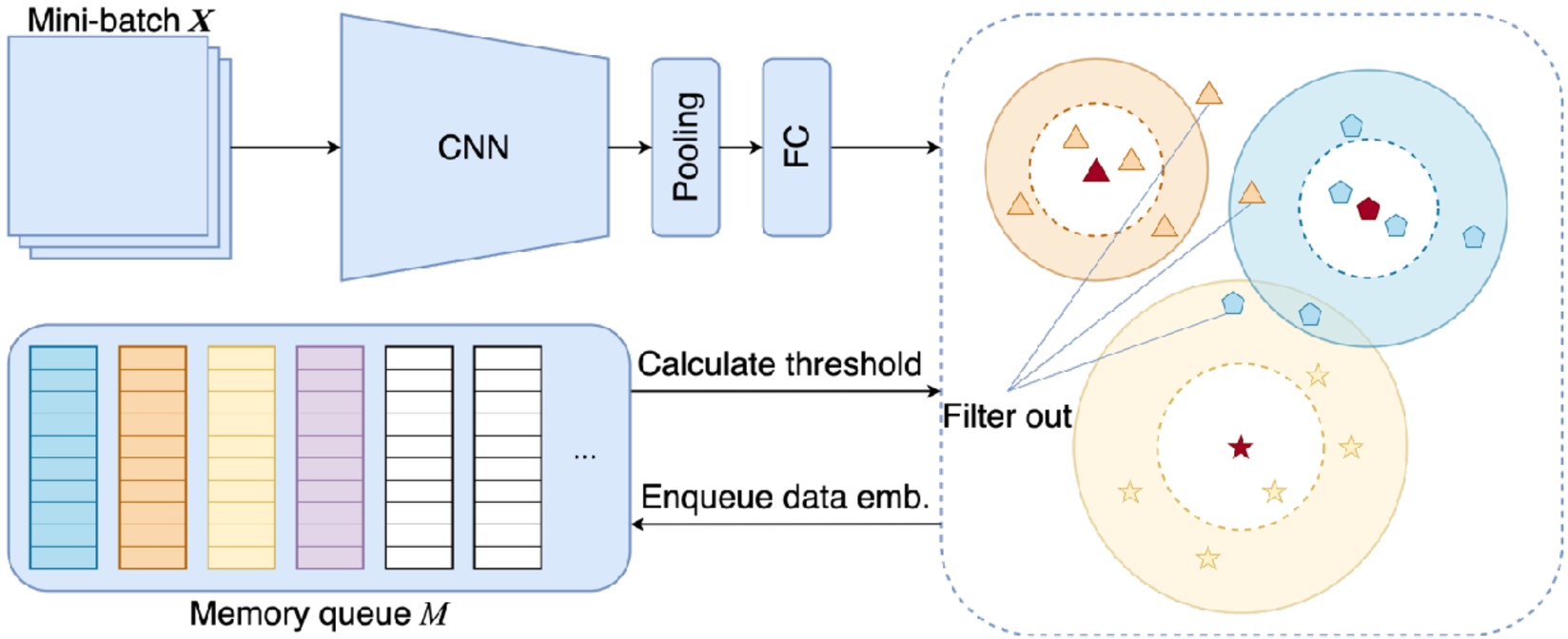
# Contributions of This Work

- Class hardness varies along learning → dynamically estimate the **class hardness**.
- Outliers and less informative samples → class-dependent **dynamic weighting** based on learned intra/inter-class relations.
- Informative **S**ample-**A**ware Proxy (Proxy-ISA).



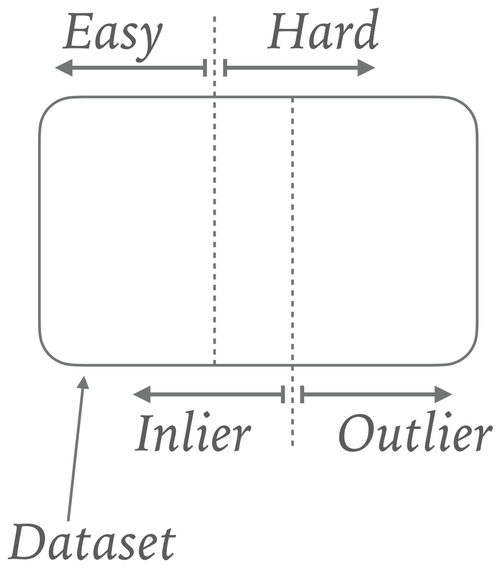
# Estimate the Class Hardness Base on Historical Embeddings

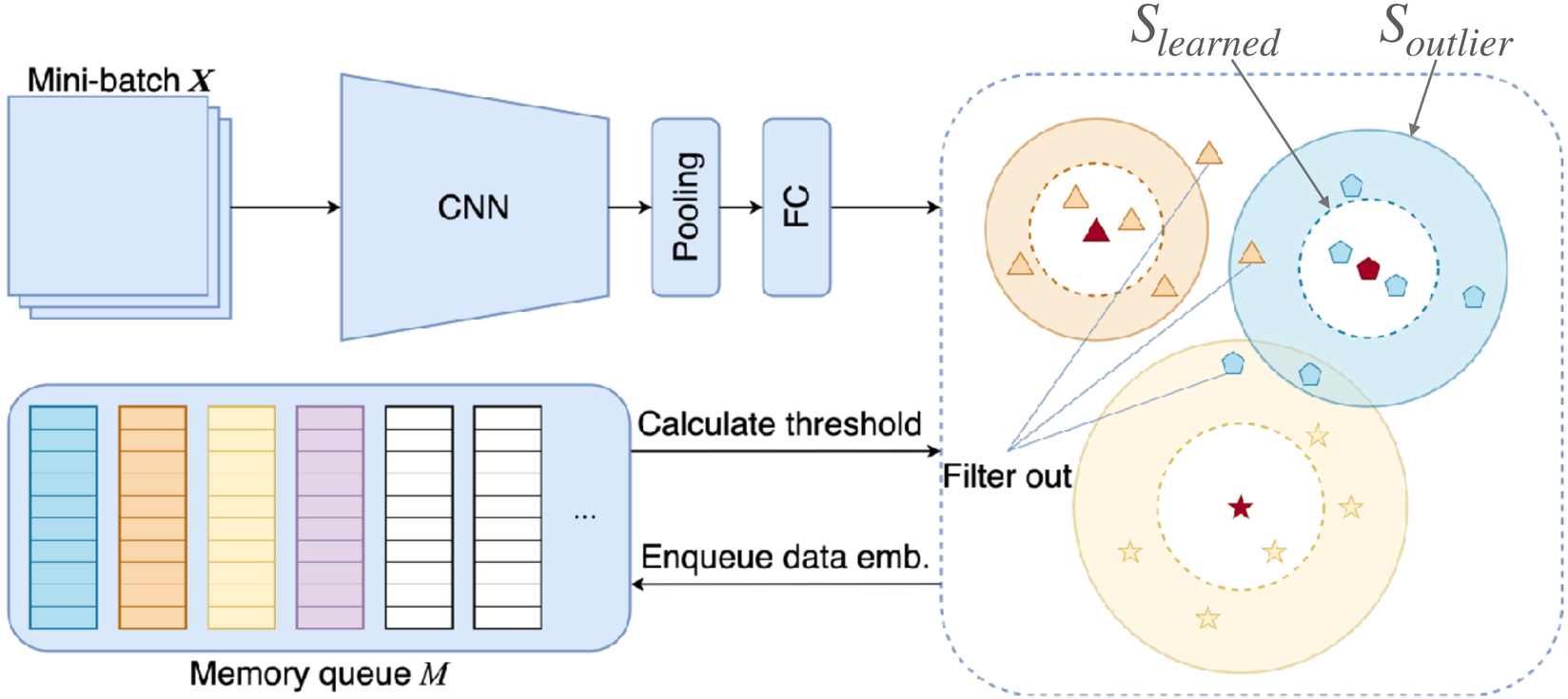
- The data learned in recent mini-batches → information about embedding space in last few iterations.
- Define total num. of embeddings from class $c$ in $M$ as $T_c$, given a hardness scaling factor $h$, the class hardness of class $c$ can be estimated by: $S_{learned}(c) = \frac{h}{T_c} \sum_{(x_i,y_i)\in M, y_i=c} S_{ic}$. ← *Threshold of less informative samples*
- After a mini-batch training, filter out outliers and enqueue the rest to $M$.



# Hardness-Related Thresholds

- $S_{learned}$: when the proxy-data similarity is larger than this threshold (too close/**easy**), regard as less informative sample.
- $S_{outlier}$: when the similarity is smaller than this threshold (too far/**hard**), regard as outlier.
- $S^{(c)}_{outlier} = S^{(c)}_{learned} - \eta_c \in (-1, S^{(c)}_{learned})$



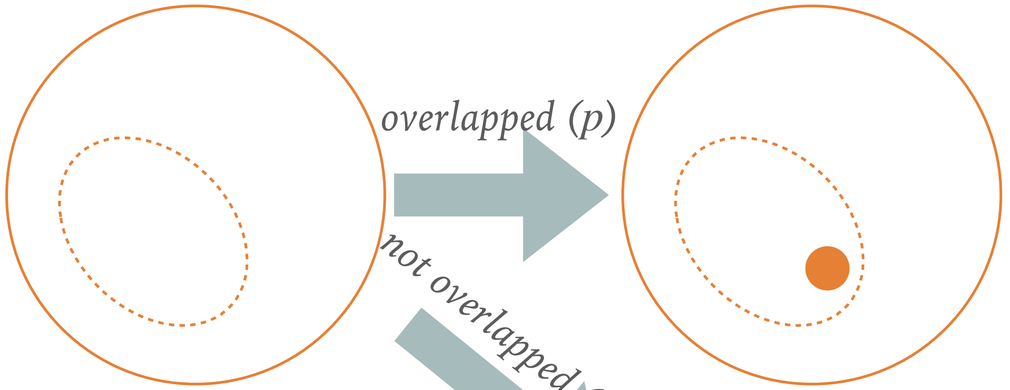



# How Can We Estimate the Learned Class ?

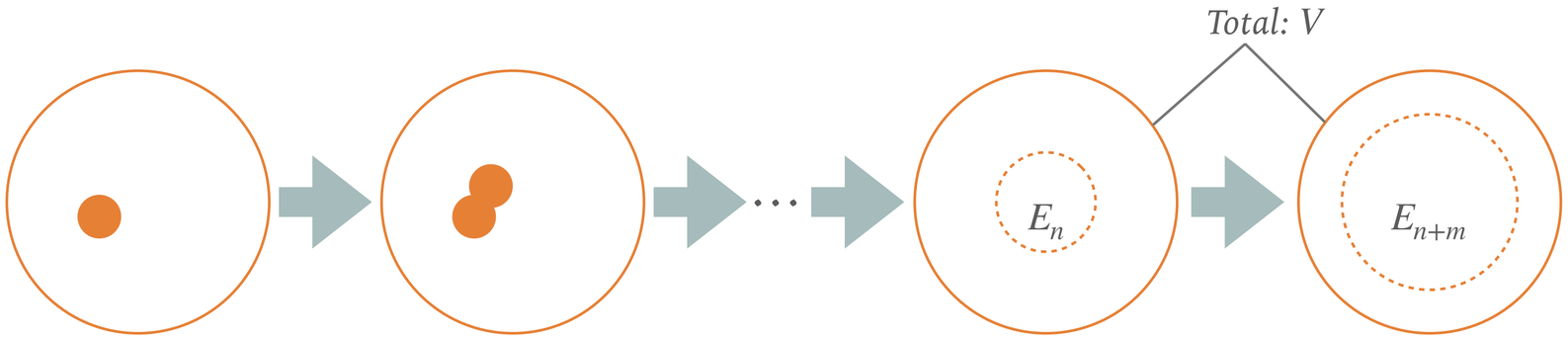
➤ Effective Number (Y. Cui+, 2019 [14])

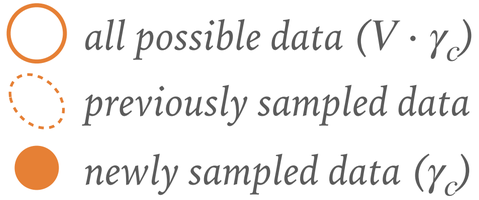
*$\gamma_c$: volumetric unit of class c*

- Volume of correctly learned data for a class while the $n$th sample of this class is selected: $E_n \cdot \gamma_c$.
- Total volume of embedding space for a class: $V \cdot \gamma_c$; Random sampling → random covering with $p = \frac{E_{n-1}}{V}$.
- $E_n = pE_{n-1} + (1-p)(E_{n-1}+1) = 1 + \frac{V-1}{V}E_{n-1}$
- Let $\beta = \frac{V-1}{V}$, $E_1 = 1$, we can induce $E_n = \frac{1-\beta^n}{1-\beta} = \sum_{i=1}^{n} \beta^{i-1}$.
- $\lim_{n\to\infty} E_n = V$.




➤ $E_n$ can be used to estimate how much knowledge about a class is learned.



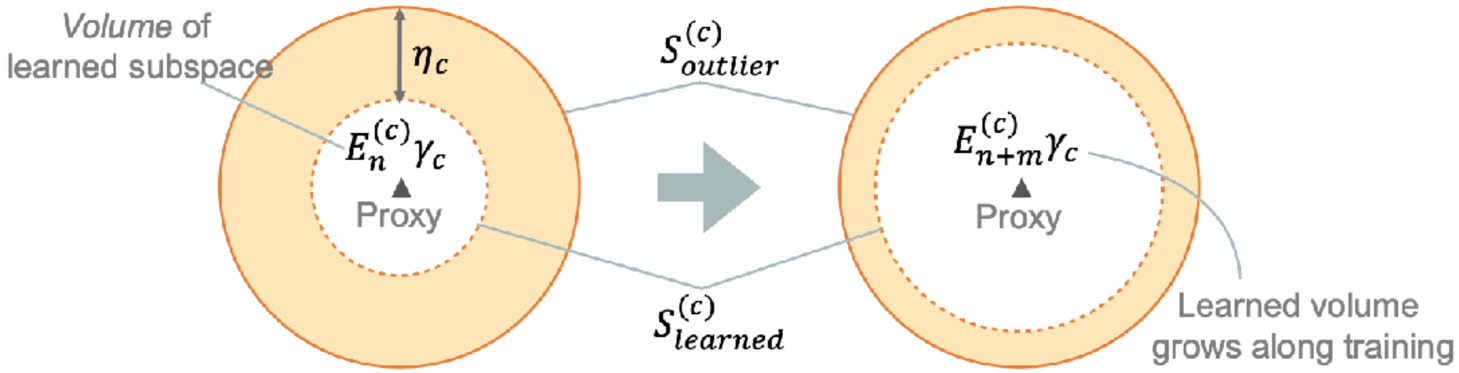



# Definition of Class-Dependent Informative Samples

- When $E_n^{(c)}$ increases, $\eta_c$ (range for informative samples) should decrease, the penalties for less informative samples and outliers should be greater.

- $\eta_c = (1 + k(1 - S_{learned}(c)))\nu_n^{(c)} + \lambda$, where $\nu_n^{(c)} = \dfrac{1}{1 + \log(1 + E_n^{(c)})}$

- $k$: **sensitivity factor**; $\lambda$: **hardness margin**



# Adaptive Weighting Factors

Dynamic weights according to different semantic states.

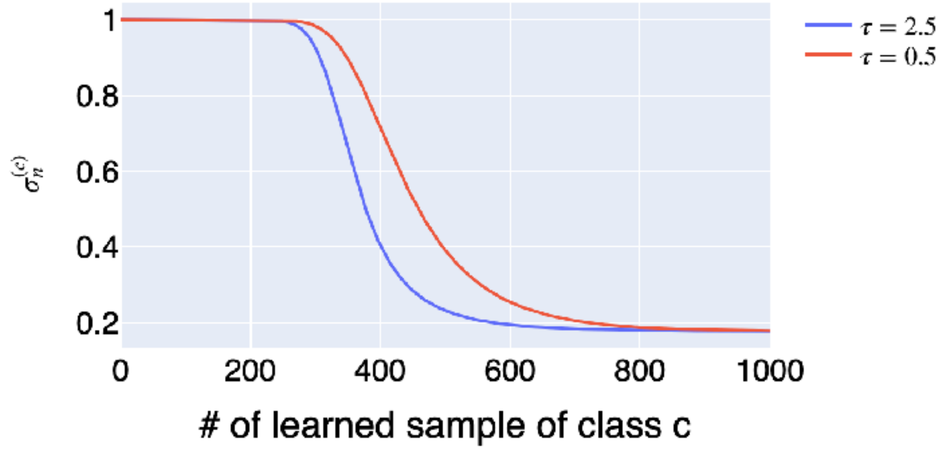
- For positive pairs:
  - $S_{ic} < S_{learned}(c) - \eta_c$, $\omega_{ic} = \sigma(c)$;
  - $S_{ic} > S_{learned}(c)$, $\omega_{ic} = \sigma(c)$;
  - Otherwise, $\omega_{ic} = 1 + \sigma(c)$.

*Informative*

- For negative pairs:
  - $S_{ic} < S_{learned}(c) - \eta_c$, $\omega_{ic} = \frac{1}{\max(1, E_n^{(c)})}$;
  - Otherwise, $\omega_{ic} = 1$.

*Informative*

$$\sigma_n^{(c)} = 1 + \frac{(1 + e^{-\tau})\left(\nu_n^{(c)} - 1\right)}{1 + e^{V - E_n^{(c)} - \tau}}$$




$$\nu_n^{(c)} = \frac{1}{1 + \log(1 + E_n^{(c)})}$$

$$\lim_{n \to \infty} \sigma_n^{(c)} = \lim_{n \to \infty} \nu_n^{(c)} = \frac{1}{1 + \log(1 + V)}$$

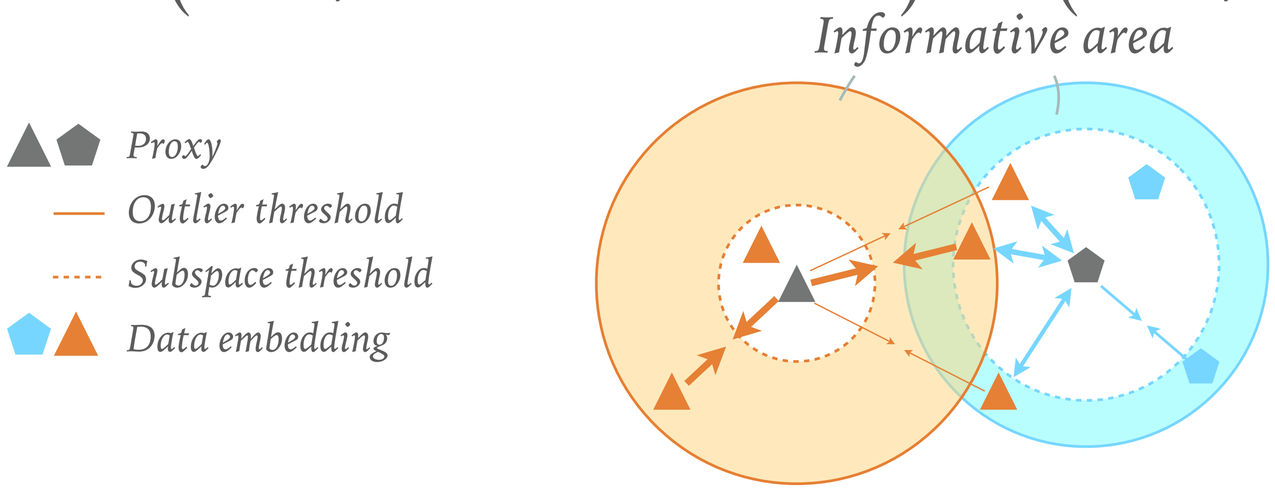
# Reform the Optimization Problem of Pair Constraints

$$\max_{\mathscr{P}_c^+} \alpha \sum_{i:y_i=c} \mathscr{P}_c^+(i)(\delta - S_{ic}) + H(\mathscr{P}_c^+) \rightarrow \max_{\mathscr{Q}_c^+} \alpha \sum_{i:y_i=c} \omega_{ic}^+ \mathscr{Q}_c^+(i)(\delta - S_{ic}) + H(\mathscr{Q}_c^+)$$

$$\max_{\mathscr{P}_c^-} \alpha \sum_{i:y_i\neq c} \mathscr{P}_c^-(i)(S_{ic} + \delta) + H(\mathscr{P}_c^-) \rightarrow \max_{\mathscr{Q}_c^-} \alpha \sum_{i:y_i\neq c} \omega_{ic}^- \mathscr{Q}_c^-(i)(S_{ic} + \delta) + H(\mathscr{Q}_c^-)$$

According to the K.K.T. condition, the optimal is

$$\log\left(1 + \sum_{i:y_i=c} \exp\{\omega_{ic}^+ \cdot \alpha(\delta - S_{ic})\}\right), \log\left(1 + \sum_{i:y_i\neq c} \exp\{\omega_{ic}^- \cdot \alpha(\delta + S_{ic})\}\right), \text{respectively.}$$

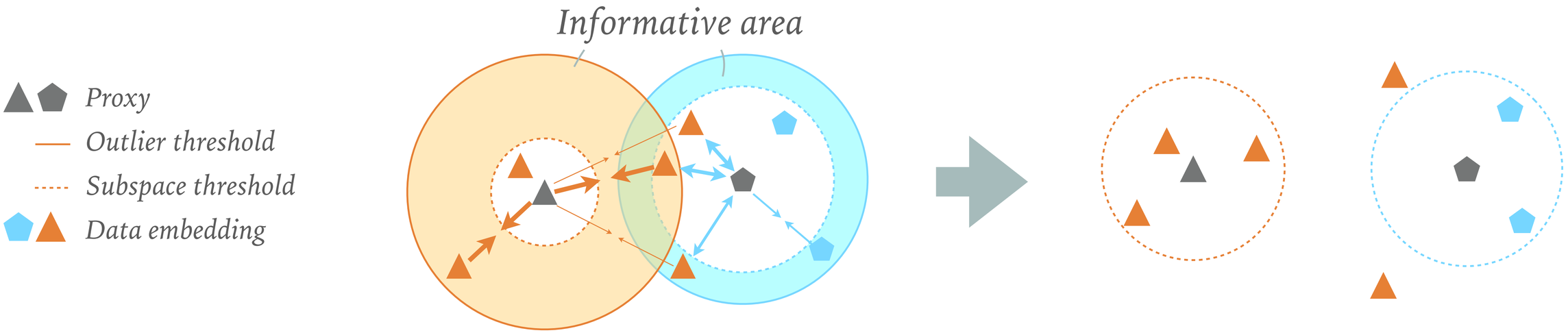


# Informative Sample-Aware Proxy (Proxy-ISA)

Formally,

$$L_{ISA} = \frac{1}{\sum_{c \in C^+} \bar{\omega}_c^+} \sum_{c \in C^+} \log \left( 1 + \sum_{i:y_i=c} \exp\{\omega_{ic}^+ \cdot \alpha(\delta - S_{ic})\} \right) + \frac{1}{\sum_c \bar{\omega}_c^-} \sum_{c=1}^{C} \log \left( 1 + \sum_{i:y_i \neq c} \exp\{\omega_{ic}^- \cdot \alpha(\delta + S_{ic})\} \right)$$

$$\frac{\partial L_{\text{ISA}}}{\partial S_{ic}} = \begin{cases} \dfrac{\omega_{ic}^+}{\sum_{c' \in C^+} \bar{\omega}_{c'}^+} \cdot \dfrac{\alpha \exp\{\omega_{ic}^+ \cdot \alpha(\delta - S_{ic})\}}{1 + \sum_{j:y_j=c} \exp\{\omega_{jc}^+ \cdot \alpha(\delta - S_{jc})\}}, & \text{if } y_i = c, \\ \dfrac{\omega_{ic}^-}{\sum_{c'} \bar{\omega}_{c'}^-} \cdot \dfrac{\alpha \exp\{\omega_{ic}^- \cdot \alpha(\delta + S_{ic})\}}{1 + \sum_{j:y_j \neq c} \exp\{\omega_{jc}^- \cdot \alpha(\delta + S_{jc})\}}, & \text{otherwise}. \end{cases}$$



# Experiments: Dataset

| Datasets | Num. of images | Num. of categories | Preview |
|---|---|---|---|
| Caltech-UCSD Birds-200-2011 (CUB) [25] | 6,033 | 200 (the first 100 for training) |  |
| Cars-196 [26] | 16,185 | 196 (the first 98 for training) |  |
| Stanford Online Products (SOP) [27] | 120,053 | 22,634 (the first 11,318 for training) |  |
| In-shop Clothes Retrieval (In-shop) [28] | 52,712 | 7,982 (the first 3,997 for training) |  |

# Experiments: Comparison to SOTA Methods

| Method | | CUB | | Cars | | SOP | | In-shop |
|---|---|---|---|---|---|---|---|---|
| | | Recall@1 | MAP@R | Recall@1 | MAP@R | Recall@1 | MAP@R | Recall@1 |
| Pair-based | Margin [6] | 63.60 | 23.09 | 81.16 | 24.21 | 70.99 | 41.82 | — |
| | MS [11] | 65.04 | 24.70 | 85.14 | 28.07 | 74.50 | 45.79 | — |
| Proxy-based | Proxy-NCA [10] | 65.01 | 23.85 | 83.56 | 25.38 | 75.89 | 47.22 | 86.92 |
| | Norm. SoftMax [23] | 65.65 | 25.25 | 83.16 | 26.00 | 75.67 | 47.13 | — |
| | SoftTriplet [16] | 66.17 | 25.61 | 84.49 | 27.08 | 76.12 | 47.35 | — |
| | CosFace [19] | 67.32 | 26.70 | 85.52 | 27.57 | 75.79 | 46.92 | — |
| | ArcFace [20] | 67.50 | 26.45 | 85.44 | 27.22 | 76.20 | 47.41 | — |
| | Proxy-Anchor [9] | 66.34 | 25.68 | 83.56 | 27.09 | 78.12 | 51.28 | 91.18 |
| | Proxy-ISA (Ours) | **68.05** | **26.79** | **86.25** | **29.29** | **78.73** | **51.52** | **92.33** |

*Training settings:*

**batch size** = 128; **embedding size** = 512;

**model**: BN-Inception [22];

**learning rate**:
1e-4 for CUB and Cars, 6e-4 for SOP and In-shop;

**Hyper-parameters**: $V$=100, $T$=55,000, $\lambda$=0.1, $\tau$=1.5

Image retrieval tasks evaluated by Recall@$K$, MAP@R;

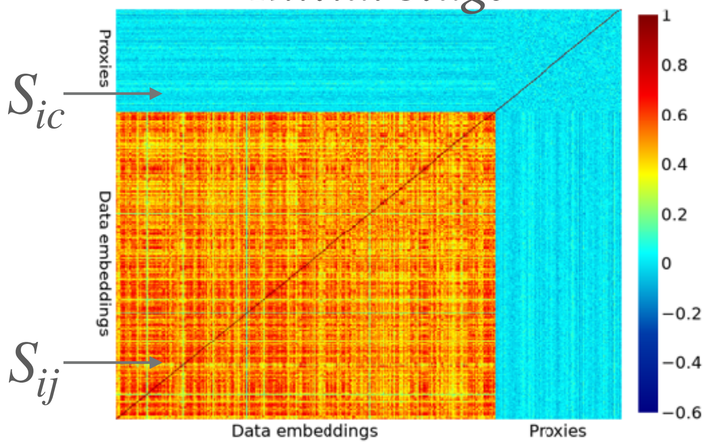
$$\text{Recall@}K\text{: } \begin{cases} 0, & \text{if none of the } K \text{ retrieval(s) is correct,} \\ 1, & \text{otherwise.} \end{cases}$$

$$\text{Mean Average Precision (MAP@R): } \frac{1}{R}\sum_{i=1}^{R} P(i)$$

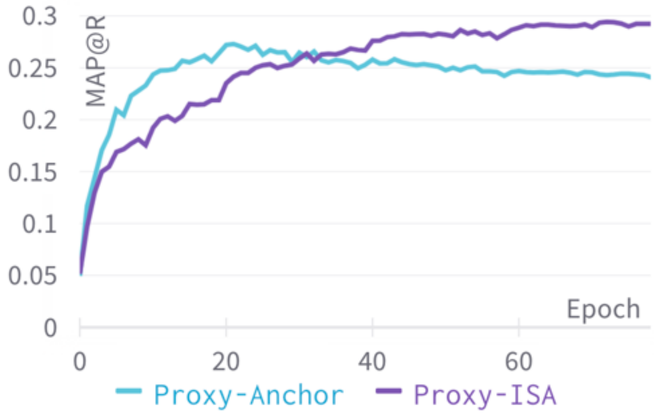
$$P(i) = \begin{cases} \text{precision@}i, & \text{if the i-th retrieval is correct,} \\ 0, & \text{otherwise.} \end{cases}$$

# Experiments: Impacts of Proxy-ISA on the Embedding Space

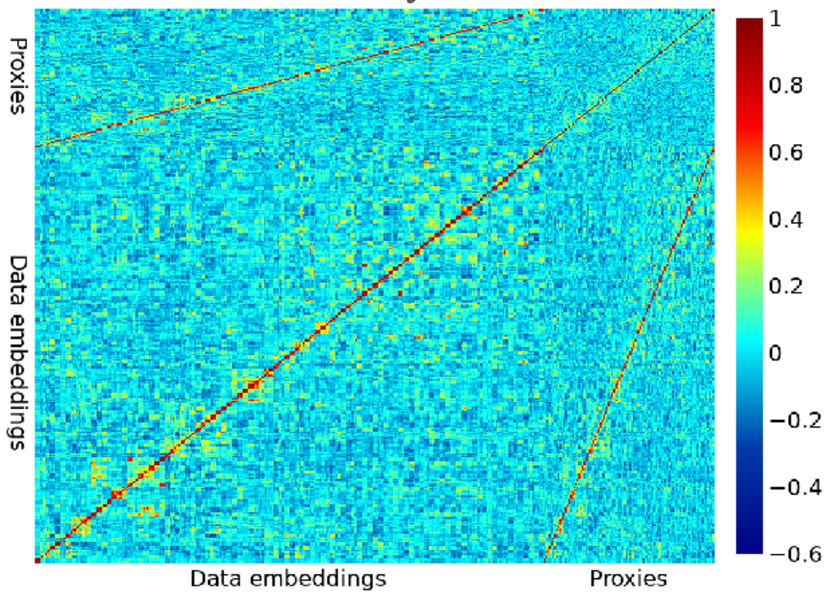
*Initial stage*




*Heat map of **cosine similarity***

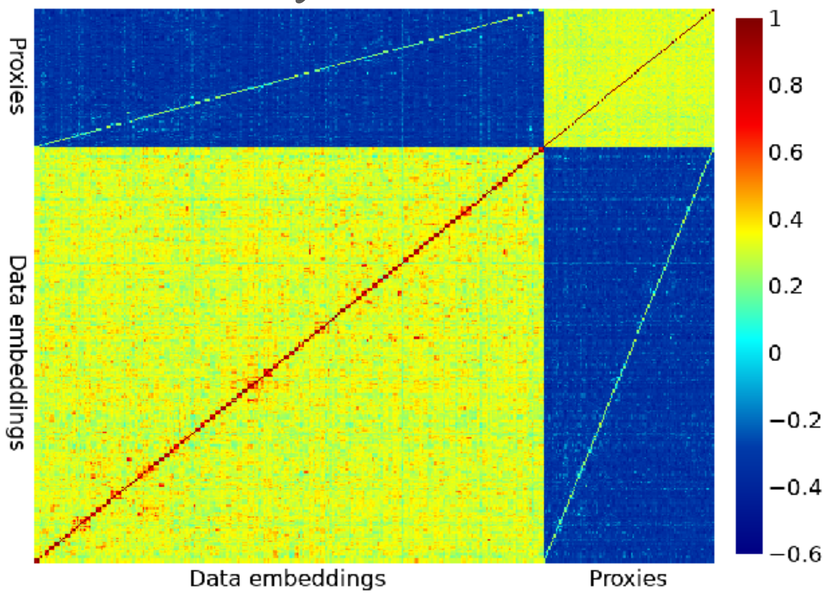



*Norm. SoftMax* [23]



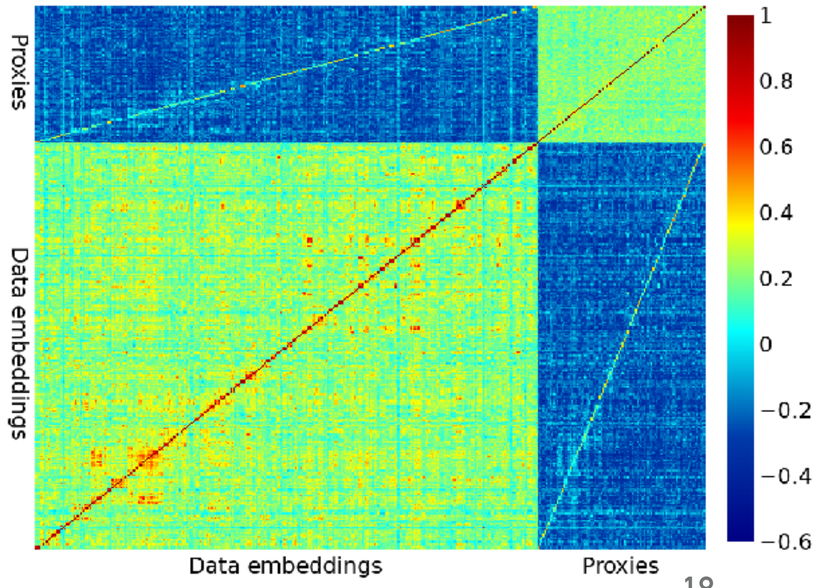

*Class hardness only*

*Proxy-Anchor* [9]




*Self + relative hardness*

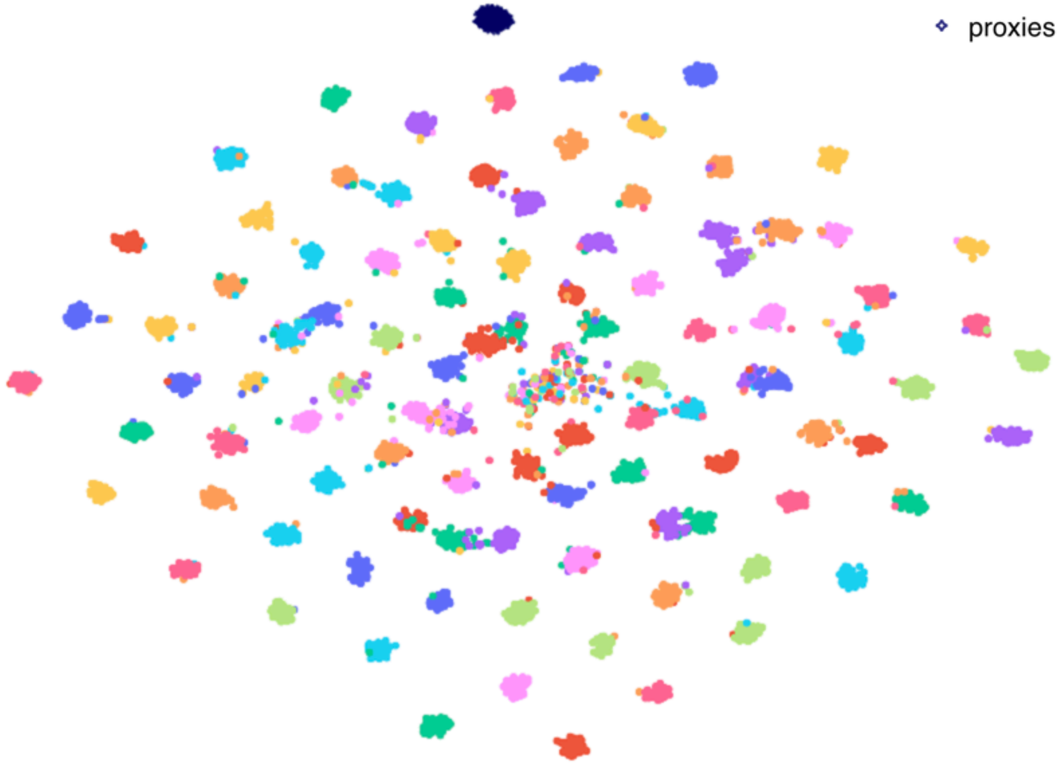
*Proxy-ISA (Ours)*



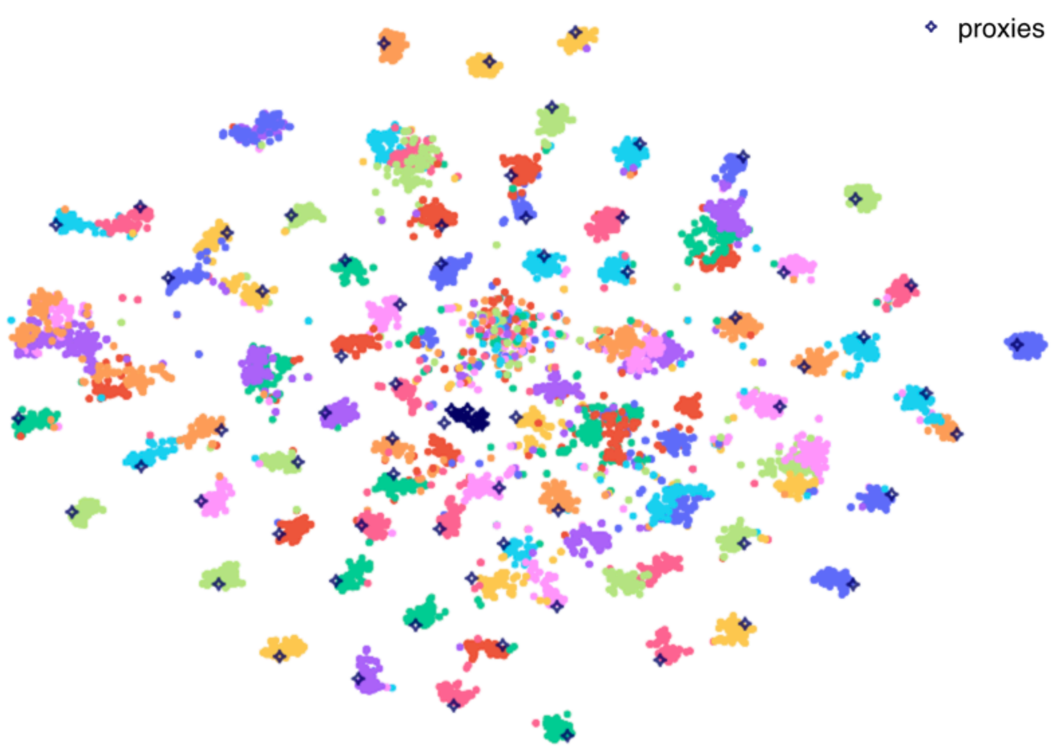

*Self + relative + class hardness*

# Experiments: Embedding Space Visualization

- t-SNE [24] visualization of Cars-196 [26] (training set)



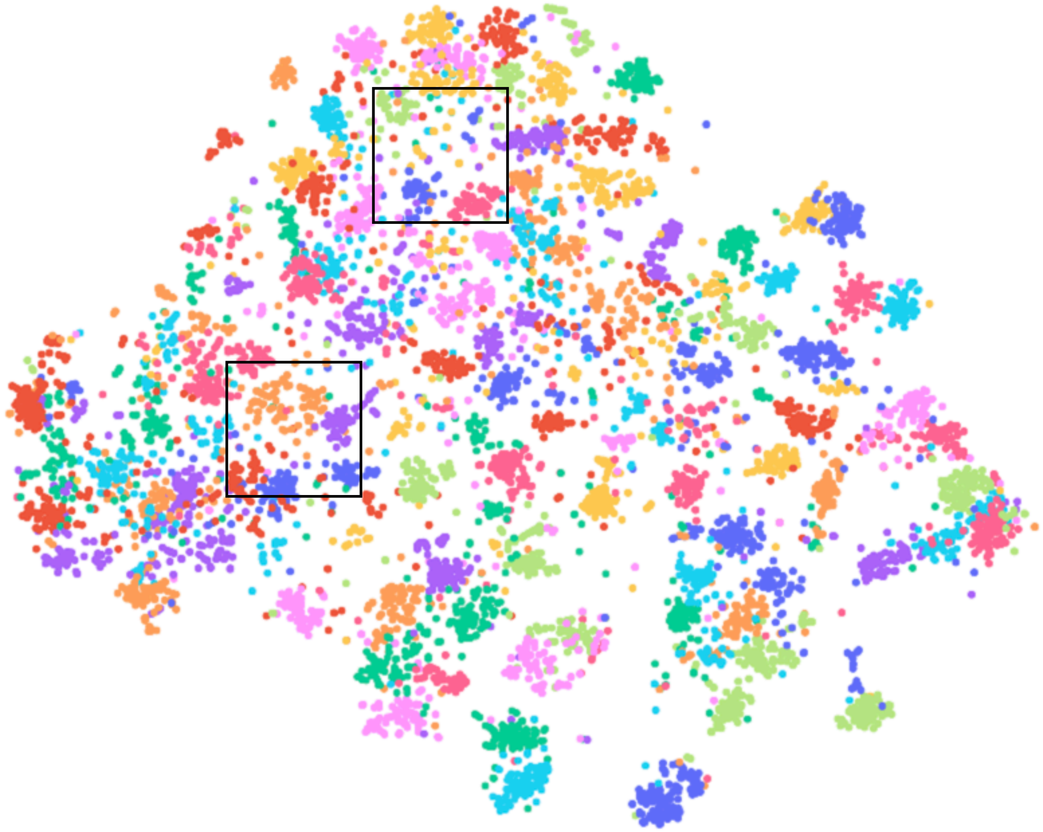

*Proxy-Anchor* [9]



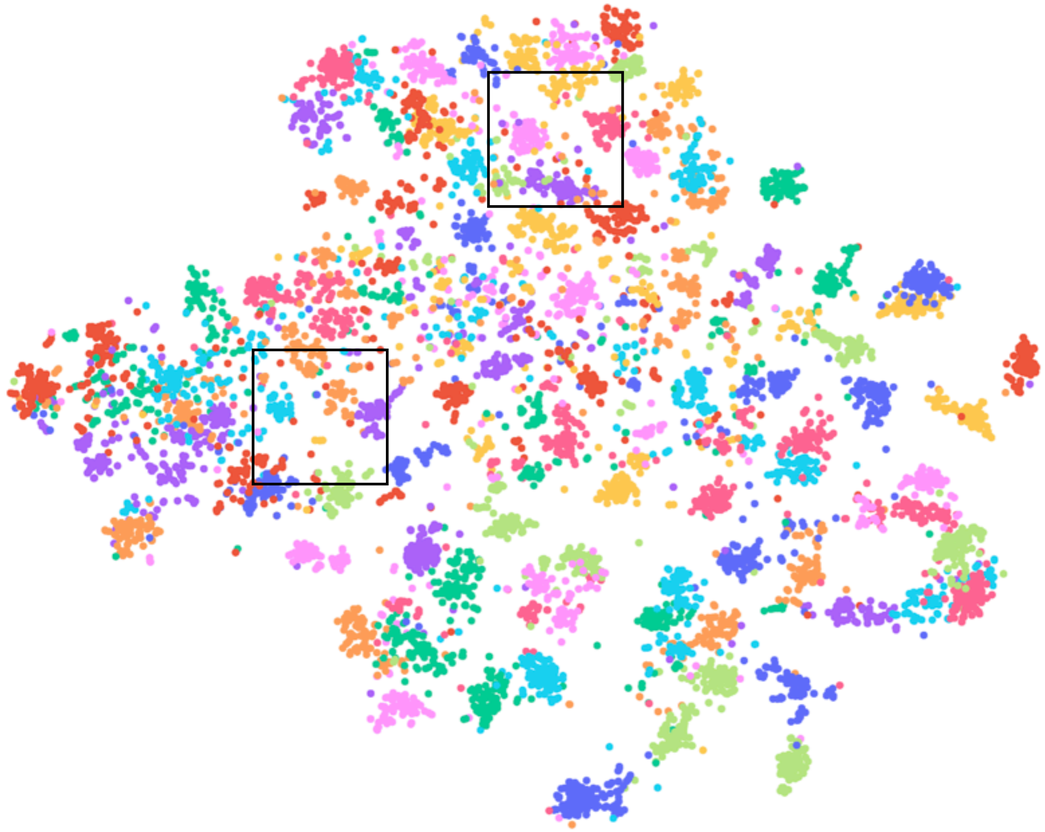

*Proxy-ISA (Ours)*

# Experiments: Embedding Space Visualization

- t-SNE [24] visualization of Cars-196 [26] (test set)



*Proxy-Anchor* [9]



*Proxy-ISA (Ours)*

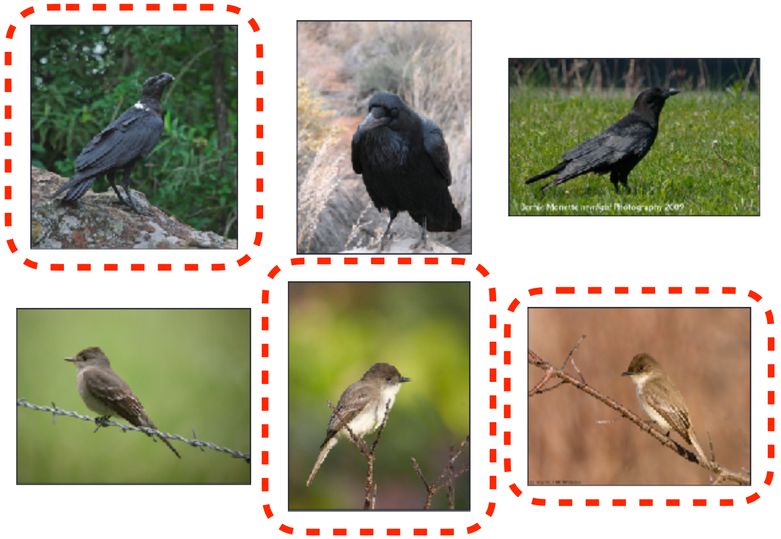
# Results of Image Retrieval

*query*

*top-3 retrievals with Proxy-Anchor* [9]

*top-3 retrievals with Proxy-ISA (Ours)*

*CUB* [25]*:*



*Cars* [26]*:*

*similar* → *dissimilar*

*failure case*

ACM Multimedia Asia 2022

# INFORMATIVE SAMPLE-AWARE PROXY FOR DEEP METRIC LEARNING

*Aoyu Li, Ikuro Sato, Kohta Ishikawa, Rei Kawakami, Rio Yokota*

aoyuli@rio.gsic.titech.ac.jp

Poster #39

Starts at 14:30

Code →

https://github.com/rioyokotalab/ProxyISA

DENSO IT LAB

# References


[1] Kishida I. et al. Empirical Study of Easy and Hard Examples in CNN training. https://arxiv.org/abs/1911.10739

[2] Chang H. et al. Active Bias: Training More Accurate Neural Network by Emphasizing High Variance Samples. NIPS 2017.

[3] Lin T. et al. Focal Loss for Dense Object Detection. ICCV 2017.

[4] Li B. et al. Gradient Harmonized Single-Stage Detector. AAAI 2019.

[5] Chopra S. et al. Learning a similarity metric discriminatively, with application to face verification. CVPR 2005.

[6] Wu C. et al. Sampling Matters in Deep Embedding Learning. ICCV 2017.

[7] Duan Y. et al. Deep Embedding Learning with Discriminative Sampling Policy. CVPR 2019.

[8] Katharopoulos A. et al. Not All Samples Are Created Equal-Deep Learning with Importance Sampling. ICML 2018.

[9] Kim S. et al. Proxy Anchor Loss for Deep Metric Learning. CVPR 2020.

[10] Movshovitz-Attias Y. et al. No Fuss Distance Metric Learning using Proxies. ICCV 2017.

[11] Wang X. et al. Multi-Similarity Loss with General Pair Weighting for Deep Metric Learning. CVPR 2019.

[12] Song H. et al. Deep Metric Learning via Lifted Structured Feature Embedding. CVPR 2016.

[13] Opitz M. et al. BIER - Boosting Independent Embeddings Robustly. ICCV 2017.

[14] Cui. Y. et al. Class-Balanced Loss Based on Effective Number of Samples. CVPR 2019.

[15] Wang X. et al. Cross-Batch Memory for Embedding Learning. CVPR 2020.

[16] Qian Q. et al. Softtriple Loss: Deep Metric Learning without Triplet Sampling. ICCV 2019.

[17] Liu C. et al. Noise-resistant Deep Metric Learning with Ranking-based Instance Selection. CVPR 2021.

[18] Hoffer E. et al. Deep metric learning using triplet network. In *Similarity-Based Pattern Recognition*, pages 84–92. Springer International Publishing, 2015.

[19] Wang H. et al. Cosface: Large margin cosine loss for deep face recognition. CVPR 2018.

[20] Deng J. et al. Arcface: Additive angular margin loss for deep face recognition. CVPR 2019.

[21] Sohn K. Improved deep metric learning with multi-class n-pair loss objective. NIPS 2016.

[22] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal co- variate shift. ICML 2015.

[23] Wang F. et al. Normface: L2 hypersphere embedding for face verification. ICM 2017.

[24] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. Journal of Machine Learning Research (2008).

[25] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J Belongie. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNSTR-2011-001. California Institute of Technology.

[26] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D Object Representations for Fine-Grained Categorization. ICCV 2013.

[27] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep Metric Learning via Lifted Structured Feature Embedding. CVPR 2016.

[28] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering Robust Clothes Recognition and Retrieval With Rich Annotations. CVPR 2016.