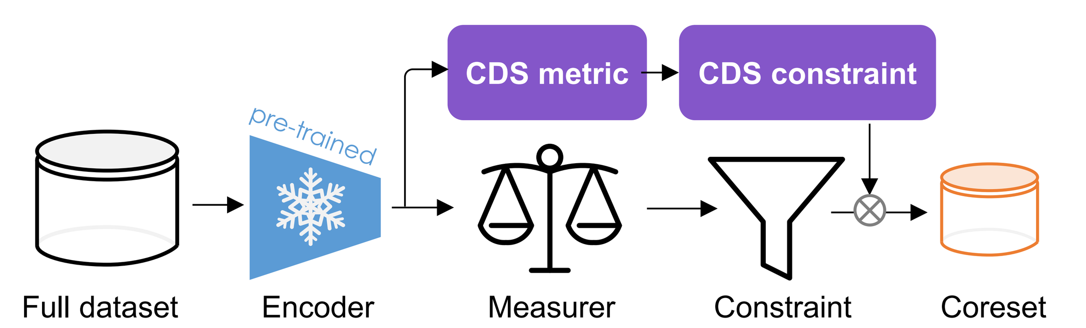
The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20-27, 2024 | VANCOUVER, CANADA

# Contributing Dimension Structure of Deep Feature for Coreset Selection


Zhijing Wan[1], Zhixiang Wang[2,3], Yuran Wang[1], Zheng Wang[1],
Hongyuan Zhu[4], and Shin'ichi Satoh[3,2]
[1]Wuhan University, [2]The University of Tokyo, [3]National Institute of Informatics, [4]A*STAR

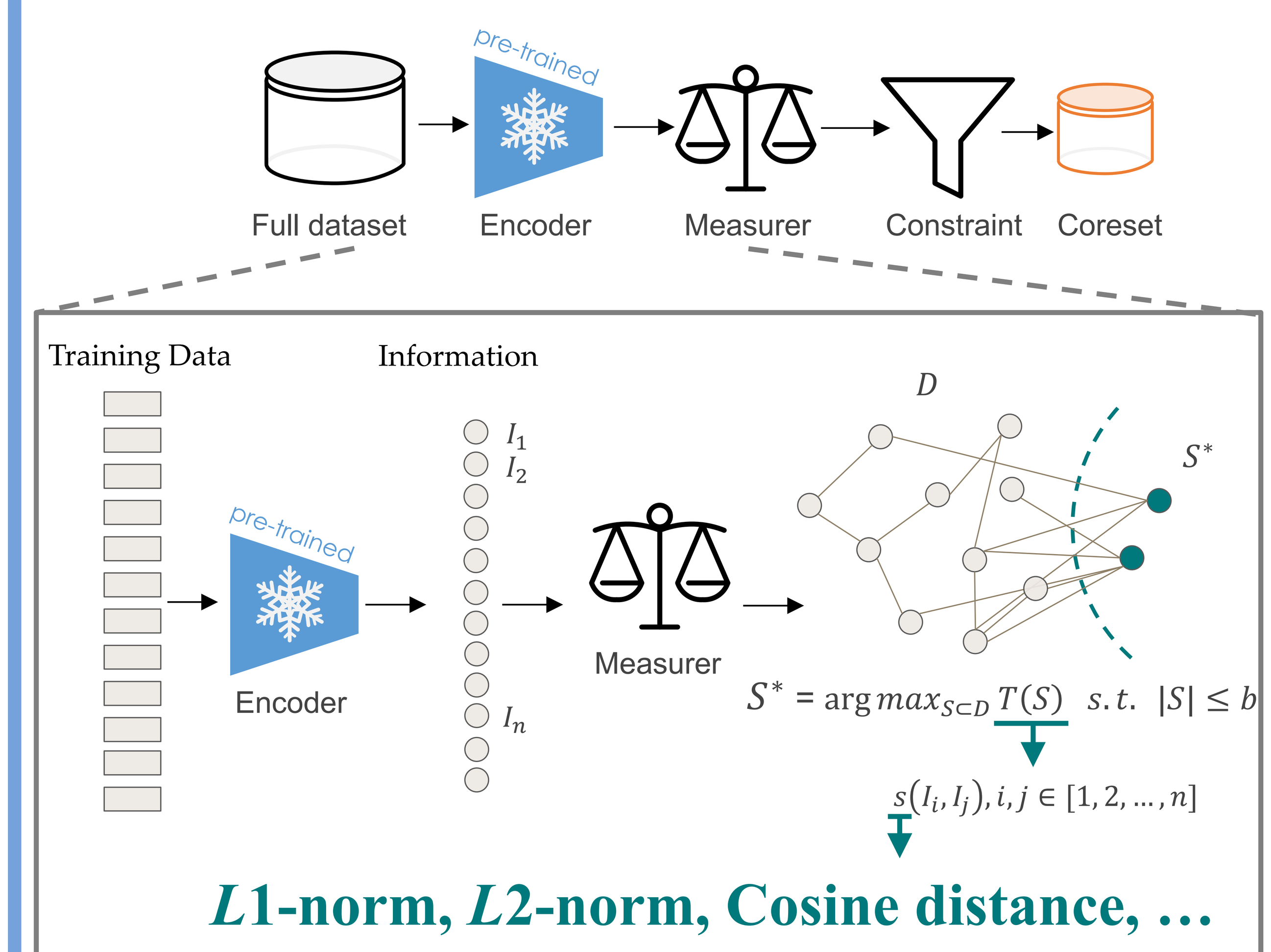

## 0. Highlights




- Propose CDS metric and constraint to successfully improve the current coreset selection pipeline;
- Propose the CDS metric to introduce the information of the Contributing Dimension Structure (CDS)
- Propose CDS constraint to enrich the diversity of CDS in the coreset

## 1. Introduction

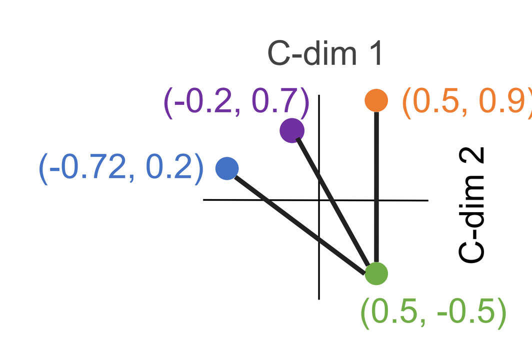
### A. Coreset selection




### B. Problems* when using similarity metrics:

Similarity metrics ignore

- the redundancy in the feature dimensions;
- disparities among the dimensions that significantly contribute to the final similarity $s$




$s(\bullet,\bullet) = s(\bullet,\bullet) = s(\bullet,\bullet)$, previous selection methods would treat them equally during the selection

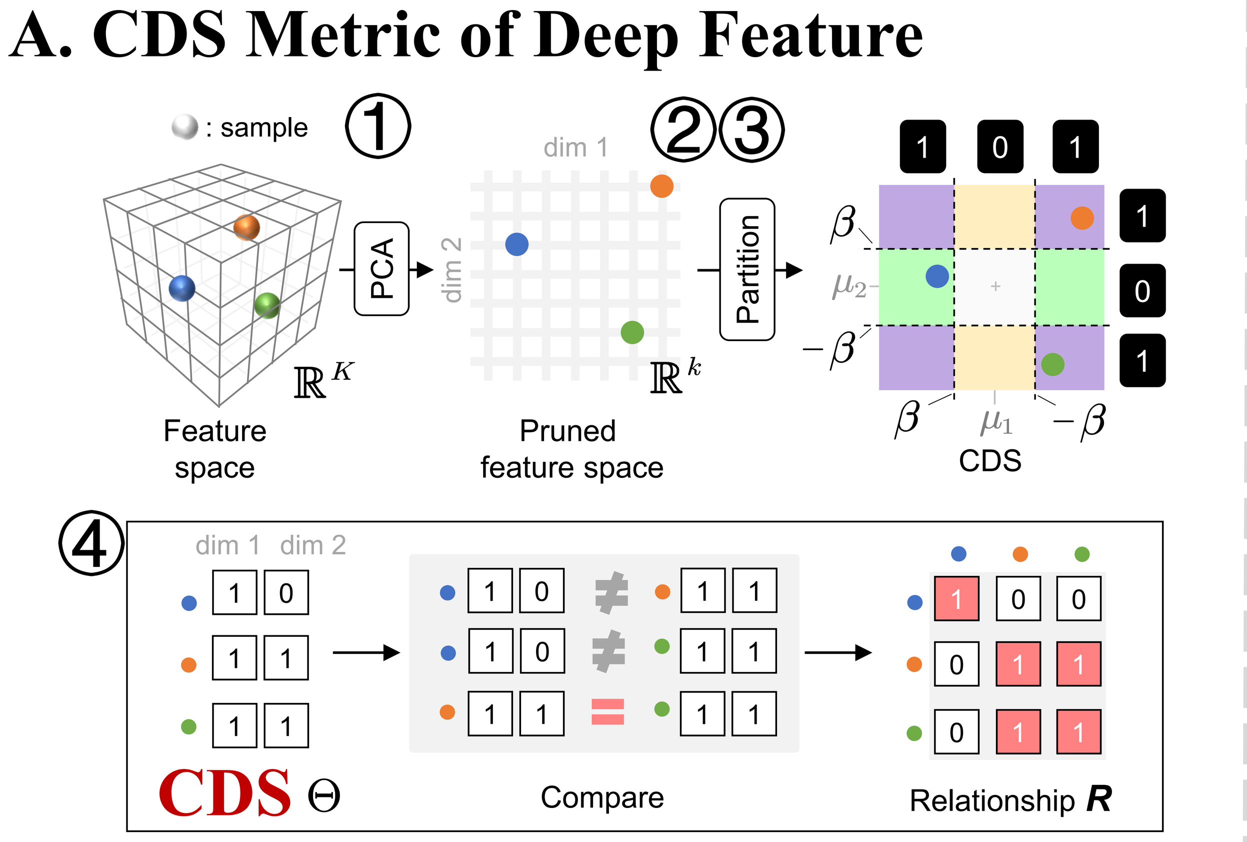
However, C-dim 1 and C-dim 2 of ●● contributes to their $s$, only C-dim 2 of ● contributes to its $s$. they have *different contributing dimensions*

Our assumption: treat them differently

*Studied specifically with feature-based selection methods using L2-norm

## 2. Methodology

### A. CDS Metric of Deep Feature




① Dimension Reduction

② Deviation from the Mean

$$\sigma = [|f_i^0 - \mu_0|, \ldots, |f_i^{k-1} - \mu_{k-1}|] \in \mathbb{R}^k,$$

Where $i \in \{0,1,\ldots,N_c - 1\}$.

③ Partition

$$\theta(x_i) = [\mathbb{I}(|f_i^0 - \mu_0|), \ldots, \mathbb{I}(|f_i^{k-1} - \mu_{k-1}|)]$$

CDS

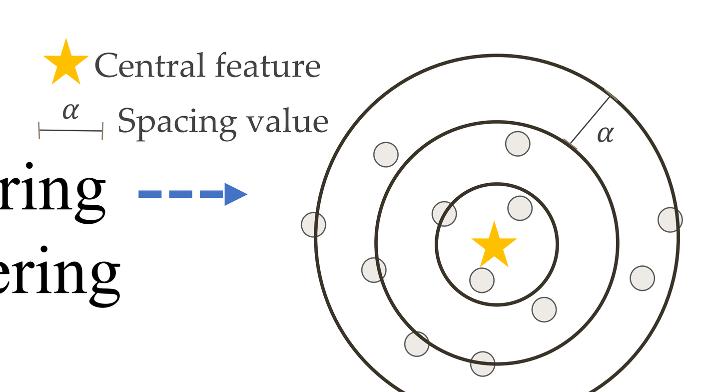
$$\mathbb{I}(\Delta f) = \begin{cases} 1, \Delta f > \beta \\ 0, \Delta f \le \beta \end{cases}$$

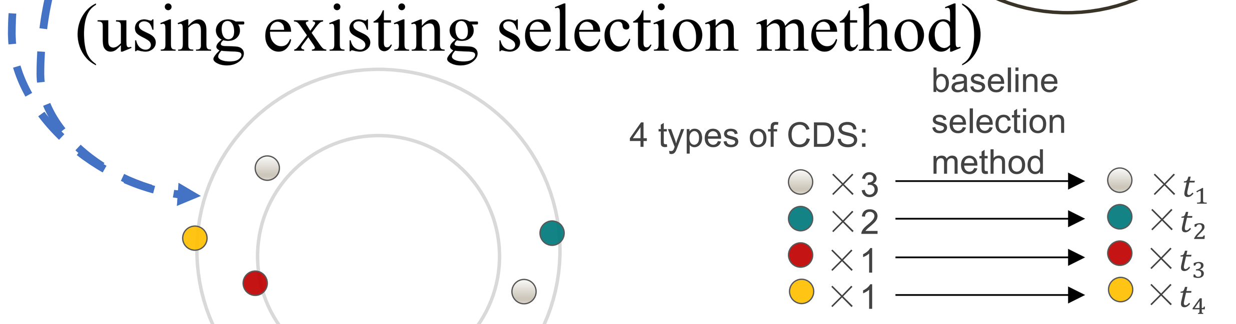
④ Comparison

### C. Coreset Selection with CDS Constraints

**Hard version**

1. $1^{st}$ stage clustering
2. $2^{nd}$ stage clustering
3. data selection (using existing selection method)




CDS Constraint: $t_1 \approx t_2 \approx t_3 \approx t_4$

**+ K-Center Greedy, Least Confidence, Moderate-DS**

**Soft version**

**Algorithm 1:** Coreset Selection with Soft CDS Constraint (Class Balanced Sampling)

**Input:** Train set: $\mathcal{D}$; data classes: $C$; budget: $b$; pre-trained model: $\theta_{pre}$; objective function: $T$; constraint function: $H$; parameter $\beta$.
**Output:** Coreset $\mathcal{S} \subset \mathcal{D}$.

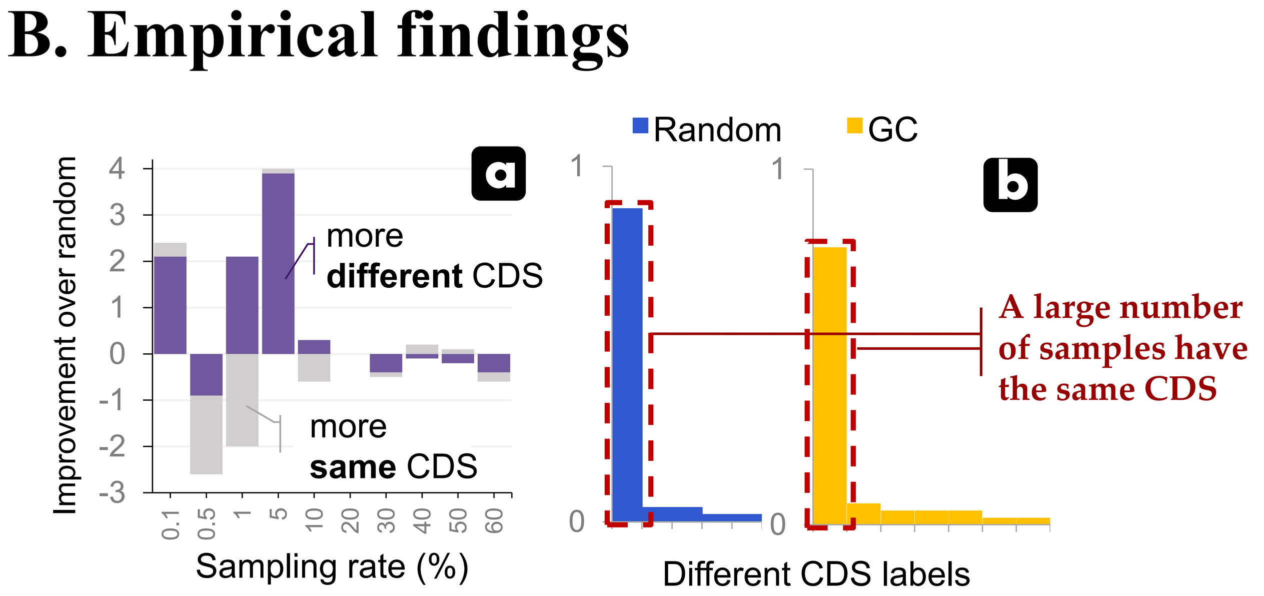
```
Initial S ← S_0
for i ← 0 to C − 1 by 1 do
    F ← M(D^i; θ_pre)
    R^i ← CDS-Metric(F, β)
    Initial S' ← S_0
    while |S'| < b do
        V ← D^i \ S'
        e ← arg max_{e∈V} T(e|S') × H(R^i)
        S' ← S' ∪ {e}
    end
    S ← S'
end
```

**+ CRAIG***

**+ GC***

*See paper for more details

### B. Empirical findings




- More data with different CDSs need to be sampled into the coreset
- The coresets selected by the existing SOTA methods are sub-optimal

Propose the **CDS Constraints** to improve SOTA selection methods

## 3. Experiments

### A. Class-balanced sampling

| Method | Sampling rates | | | | | |
|---|---|---|---|---|---|---|
| | 0.1% | 0.5% | 1% | 5% | 10% | 20% |
| Random | 18.9±0.2 | 29.5±0.4 | 39.3±1.5 | 62.4±1.7 | 74.7±1.9 | 86.9±0.3 |
| KCG | 18.7±2.9 | 27.4±1.0 | 31.6±2.1 | 53.5±2.9 | 73.2±1.3 | 86.9±0.4 |
| Forgetting | 21.8±1.7 | 29.2±0.7 | 35.0±1.1 | 50.7±1.7 | 66.8±2.5 | 86.0±1.2 |
| LC | 14.8±2.4 | 19.6±0.8 | 20.9±0.4 | 37.4±1.9 | 56.0±2.0 | 83.4±1.1 |
| CRAIG | 21.1±2.4 | 27.2±1.0 | 31.5±1.5 | 45.0±2.9 | 58.9±3.6 | 79.7±3.5 |
| Cal | 20.8±2.8 | 32.0±1.9 | 39.1±3.2 | 60.7±0.8 | 72.2±1.5 | 79.9±0.5 |
| Glister | 19.5±2.1 | 29.7±1.1 | 33.2±1.1 | 47.1±2.6 | 65.7±1.7 | 83.4±1.7 |
| GC | 22.9±1.4 | 34.0±1.3 | 42.0±3.0 | 66.2±1.0 | 75.6±1.4 | 84.3±0.4 |
| M-DS | 21.0±3.0 | 31.8±1.2 | 37.7±1.4 | 63.4±2.2 | 78.0±1.3 | 87.9±0.5 |
| GC+Ours | 24.6±1.7 | 36.4±1.0 | 43.1±1.8 | 67.1±0.6 | 76.9±0.2 | 85.2±0.6 |
| Δ | 1.7 ↑ | 2.4 ↑ | 1.1 ↑ | 0.9 ↑ | 1.3 ↑ | 0.9 ↑ |
| M-DS+Ours | 22.0±2.0 | 33.0±1.3 | 40.7±1.0 | 64.9±0.8 | 79.6±0.4 | 87.9±0.2 |
| Δ | 1.0 ↑ | 1.2 ↑ | 3.0 ↑ | 1.5 ↑ | 1.6 ↑ | 0.0 ↑ |

Table 1: Comparison on the class-balanced sampling setting. We train randomly initialized ResNet-18 on coresets of CIFAR-10 selected by different methods and then test them on the test set of CIFAR-10. Green emphasizes the best performance at each sampling rate. Δ denotes the improvement of baseline+Ours over baseline.

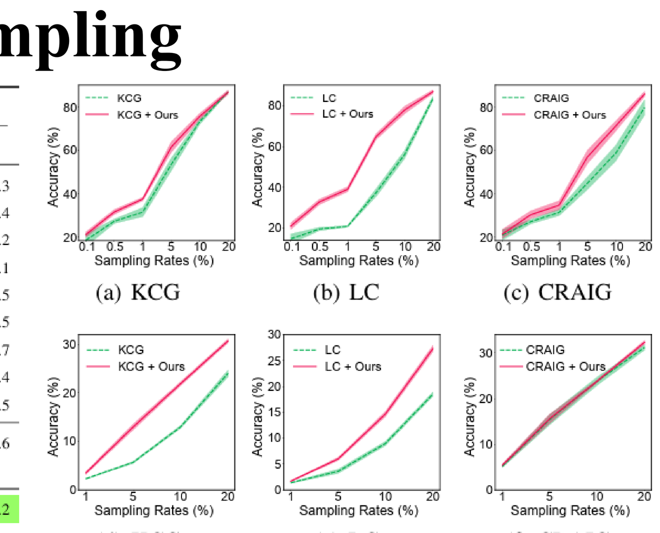



Figure 5: Performance improvement over baselines. We improve current methods with our proposed CDS metric and constraint. We compare the improved versions with respective baselines on CIFAR-10 (a–c) and TinyImageNet (d–f) under the class-balanced sampling setting. The improved versions consistently outperform baselines, suggesting that increasing the diversity of CDS in the coreset can universally enhance existing coreset selection methods.

### B. Ablation and Parameter Studies

| | dim reduction | partition | CDS-r | cons-traint | |
|---|---|---|---|---|---|
| (v1) | ✗ | ✗ | ✗ | ✗ | 34.0±1.3 |
| (v2) | ✗ | ✗ | ✗ | ✓ | 32.8±0.7 |
| (v3) | ✓ | ✗ | ✗ | ✓ | 34.3±2.5 |
| (v4) | ✓ | ✓ | ✗ | ✓ | 33.5±1.1 |
| full | ✓ | ✓ | ✓ | ✓ | 36.4±1.0 |

Table 2: Ablation study on 0.5% of the CIFAR-10

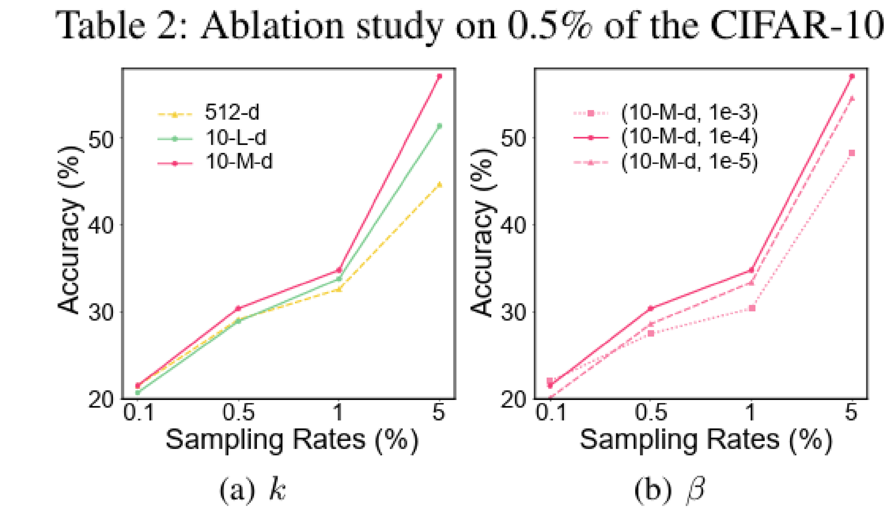
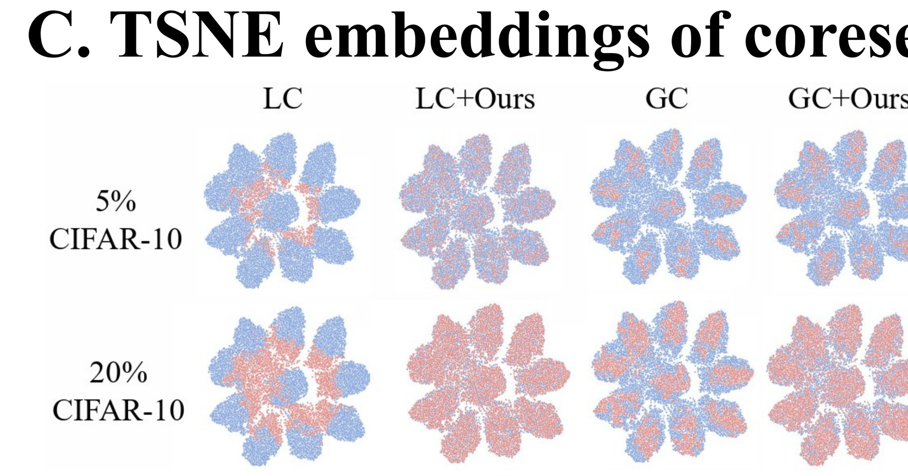



Figure 7: Parameter analysis. It shows that our method achieves the best improvement compared to the baseline method (CRAIG) when $K$ =10-M-D and $\beta$ =1e-4.

### C. TSNE embeddings of coreset




Selected data, unselected data


**Contact information: wanzjwhu@whu.edu.cn**