# CSOR: Coreset Selection for Object Re-identification via Class Pruning


Minyoung Oh and Jae-Young Sim

Graduate School of Artificial Intelligence, Ulsan National Institute of Science and Technology (UNIST)


Linkedin

## Coreset Selection (CS)

- **Concept**: Coreset selection aims **to reduce dataset size** by selecting informative samples from original dataset while **minimizing performance degradation**.

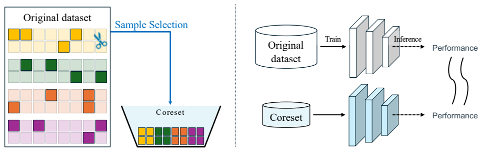


Fig 1. Basic concept of coreset selection.

## What makes CS for ReID different from Classification?

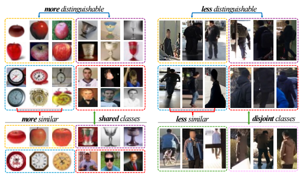


Fig 2. Comparison of (*left*) Classification and (*right*) ReID datasets.

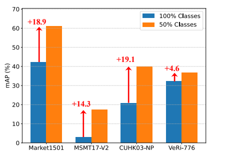


Fig 3. Random class pruning enhances ReID performance.

- **Observation 1 (*O1*)**
  : ***Coarse* vs. *Fine-grained*** characteristics across classes.
  → *O1* states ***intra-class diversity*** is important to learn class-discriminative feature.
  → To capture Intra-class diversity, enough budget per class is required.
- **Observation 2 (*O2*)**
  : ***Shared* vs. *Disjoint*** label space across train and test sets.
  → *O2* provides a feasible solution: ***discard classes*** and ***assign more budgets per class.***
  → Fig 3: ***Random class pruning*** allows more budget per class with better ReID performance.

## Extension CS to ReID datasets: *Finding Optimal Classes*

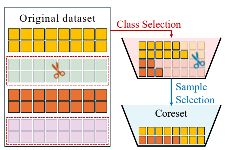


Fig 4. Concept of coreset selection for ReID

- **New Formulation of CS for ReID**
  : Find not only ***optimal samples***
  but also ***optimal classes to preserve***.

$$\widehat{\mathcal{D}}^*(\widehat{\mathcal{C}}^*) = \arg\min_{\widehat{\mathcal{D}}(\hat{\mathcal{C}})} \mathbb{E}_{\substack{x_i \sim P_{\mathcal{Q}} \\ x_j \sim P_{\mathcal{G}}}} [d(x_i, x_j; \phi_{\widehat{\mathcal{D}}(\hat{\mathcal{C}})}) | y_i = y_j],$$

$$\text{s.t.} \quad \hat{\mathcal{C}} \subseteq \mathcal{C}, \widehat{\mathcal{D}}(\hat{\mathcal{C}}) \subset \mathcal{D}, \text{ and } |\widehat{\mathcal{D}}(\hat{\mathcal{C}})| \leq B.$$

## Proposed Method: 2-Stage Framework (DCP and CPS)

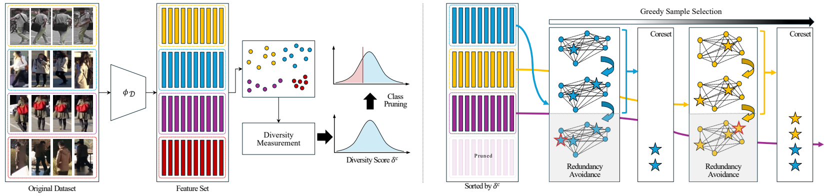


Fig 5. Overview of the proposed framework: (*left*) Diversity-driven Class Pruning, and (*right*) Coverage-Prioritized Sampling.

- **[Stage 1] Diversity-driven Class Pruning (DCP)**

  **- Core Motivation**
  : ***low-diversity*** classes are redundant and less-informative.

  **- Diversity Measurement**

$$\delta^c = \mathrm{Tr}(\mathrm{Cov}(\mathcal{F}^c))$$

  : Bottom $\alpha$ percentage of diversity score ***classes are removed.***
  : This allows remaining classes to have more samples.
  → Better ***intra-class diversity*** per class.

Fig 5. High-diversity classes (*left, Preserved*) and Low-diversity classes (*right, Pruned*).

- **[Stage 2] Coverage-Prioritized Sampling (CPS)**

  **- Core Motivation**
  : ***reducing redundancy within class*** yields better intra-class diversity.

  **- Facility Location**

$$\rho(\widehat{\mathcal{F}^c}) = \frac{1}{|\mathcal{F}^c|} \sum_{f_j \in \mathcal{F}^c} \max_{f_i \in \widehat{\mathcal{F}^c}} s_{ij} \xrightarrow{\textit{Marginal gain}} \Delta(f_k | \widehat{\mathcal{F}^c}) = \frac{1}{|\mathcal{F}^c|} \sum_{f_j \in \mathcal{F}^c} (0, s_{kj} - \max_{f_i \in \widehat{\mathcal{F}^c}} s_{ij})$$

  : For each class, iteratively select samples that ***maximize coverage gain*** with ***greedy*** manner.
  : When the gain becomes **smaller than threshold $\tau$**, ***sampling is terminated*** for that class to ***avoid redundancy***.


Fig 6. Visualization of selected (*diverse*) and pruned (*redundant*) samples for each class by CPS.

## Experimental Results

| Datasets | Market1501 | | | | MSMT17-V2 | | | | CUHK03-NP | | | | VeRi-776 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Budget | 500 | | 1,000 | | 500 | | 1,000 | | 500 | | 1,000 | | 500 | | 1,000 | |
| Ratio (%) | 3.8 | | 7.7 | | 1.5 | | 3.1 | | 6.8 | | 13.6 | | 1.3 | | 2.6 | |
| Metric | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 |
| Whole | 87.2 | 94.3 | 87.2 | 94.3 | 60.6 | 81.2 | 60.6 | 81.2 | 75.7 | 78.0 | 75.7 | 78.0 | 79.1 | 96.7 | 79.1 | 96.7 |
| Uniform | 14.7 | 31.7 | 42.3 | 63.8 | 2.0 | 6.2 | 3.1 | 9.2 | 2.7 | 2.0 | 20.9 | 19.3 | 13.0 | 33.1 | 32.2 | 57.4 |
| Random | 37.0 | 58.3 | 54.5 | 73.3 | 7.6 | 18.7 | 15.4 | 32.1 | 15.5 | 13.4 | 35.3 | 36.3 | 24.9 | 52.1 | 34.8 | 64.2 |
| Forgetting | 30.5 | 52.2 | 49.1 | 71.6 | 2.8 | 9.4 | 4.8 | 13.9 | 12.0 | 10.8 | 29.7 | 28.7 | 10.9 | 30.0 | 23.0 | 52.8 |
| AUM | 49.2 | 71.8 | 57.4 | 78.0 | 13.2 | 32.5 | 17.4 | 39.5 | 24.6 | 22.9 | 34.4 | 34.6 | 23.2 | 56.8 | 29.9 | 65.6 |
| EL2N | 46.2 | 70.2 | 59.4 | 80.4 | 4.5 | 13.0 | 3.9 | 9.8 | 16.8 | 14.6 | 31.1 | 31.2 | 19.0 | 52.3 | 26.5 | 62.7 |
| GraNd | 38.1 | 61.0 | 55.2 | 75.7 | 8.4 | 22.1 | 15.2 | 34.1 | 11.8 | 10.4 | 28.3 | 26.9 | 22.2 | 57.3 | 26.4 | 60.8 |
| Moderate | 36.3 | 57.8 | 55.8 | 75.1 | 8.3 | 20.2 | 15.5 | 33.6 | 16.4 | 15.1 | 35.9 | 34.5 | 21.4 | 49.6 | 33.8 | 62.5 |
| CCS | 43.0 | 64.2 | 57.2 | 76.3 | 10.3 | 24.7 | 16.6 | 35.1 | 17.8 | 15.4 | 35.6 | 35.8 | 26.7 | 56.0 | 35.2 | 67.3 |
| **Proposed** | **56.2** | **76.5** | **67.1** | **83.8** | **15.9** | **34.6** | **26.2** | **50.2** | **30.3** | **30.6** | **43.1** | **44.6** | **33.7** | **60.7** | **41.9** | **68.0** |

Tab 1. Performance comparison with state-of-the-art coreset selection methods on 3 person ReID datasets and 1 vehicle ReID dataset.

| Datasets | Market1501 | | | | | |
|---|---|---|---|---|---|---|
| Budget | 500 | | | 1,000 | | |
| Ratio (%) | 3.8 | | | 7.7 | | |
| Method | # Class | SPC | MMD | # Class | SPC | MMD |
| Uniform | 751 | 1.33 | 2.39 | 751 | 2.00 | 1.27 |
| Random | 514 | 1.95 | 1.87 | 590 | 2.54 | 1.49 |
| Forgetting | 451 | 2.22 | 2.03 | 549 | 2.73 | 1.69 |
| AUM | 144 | 6.94 | 1.10 | 191 | 7.85 | 1.00 |
| EL2N | 208 | 4.81 | 0.71 | 280 | 5.36 | 0.70 |
| GraNd | 411 | 2.43 | 1.78 | 505 | 2.97 | 1.55 |
| Moderate | 488 | 2.05 | 1.79 | 579 | 2.59 | 1.56 |
| CCS | 356 | 2.81 | 1.73 | 416 | 3.61 | 1.45 |
| **Proposed** | **167** | **5.99** | **0.28** | **247** | **6.07** | **0.29** |

Tab 2. Statistics of coresets constructed by different methods, SPC: Samples Per Class.

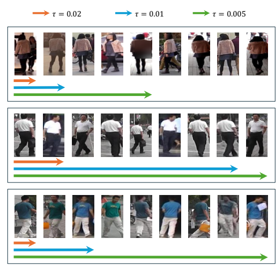


Fig 7. Visualization of selected sample by CPS with different values of $\tau$.

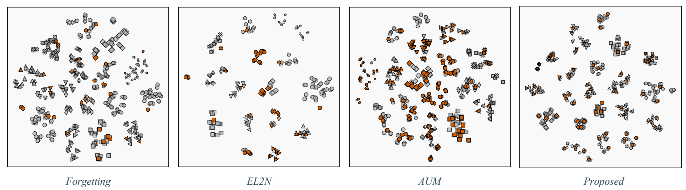


Fig 8. T-SNE visualizations of most frequently selected 15 identities by each method on Market1501 dataset under 1,000 image budget. Orange and gray color indicate selected and unselected samples, respectively.