SCoRe: Submodular Combinatorial Representation Learning

Anay Majee [1] **Suraj Kothawade** [2] **Krishnateja Killamsetty** [3] **Rishabh Iyer** [1]

Abstract

In this paper we introduce the **SCoRe**[1] (**S**ubmodular **Co**mbinatorial **Re**presentation Learning) framework, a novel approach in representation learning that addresses inter-class bias and intra-class variance. SCoRe provides a new combinatorial viewpoint to representation learning, by introducing a family of loss functions based on set-based submodular information measures. We develop two novel combinatorial formulations for loss functions, using the *Total Information* and *Total Correlation*, that naturally minimize intra-class variance and inter-class bias. Several commonly used metric/contrastive learning loss functions like supervised contrastive loss, orthogonal projection loss, and N-pairs loss, are all instances of SCoRe, thereby underlining the versatility and applicability of SCoRe in a broad spectrum of learning scenarios. Novel objectives in SCoRe naturally model class-imbalance with up to 7.6% improvement in classification on CIFAR-10-LT, CIFAR-100-LT, MedMNIST, 2.1% on ImageNet-LT, and 19.4% in object detection on IDD and LVIS (v1.0), demonstrating its effectiveness over existing approaches.

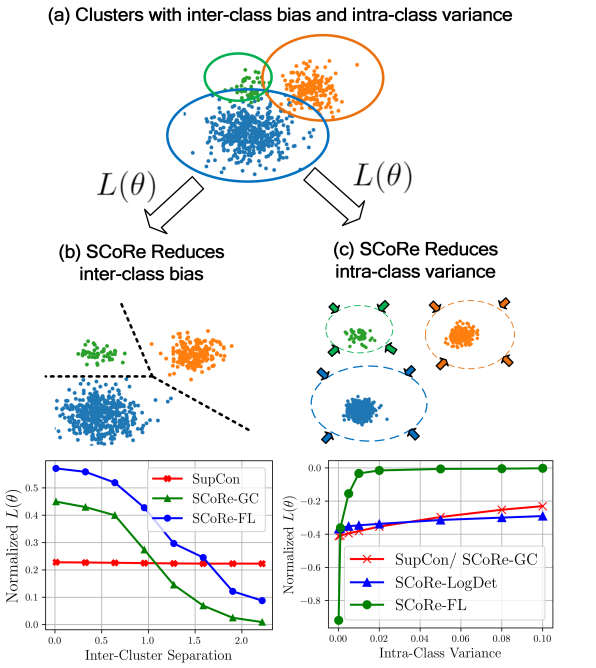


Figure 1. **Objectives in SCoRe are resilient to inter-class bias and intra-class variance in long-tail settings**. Applying $L(\theta)$ to (a) reduces inter-class bias by promoting inter-cluster separation in (b) while reducing intra-class variance in (c) by inducing intra-cluster compactness.

1. Introduction

Visual Object Recognition in real-world scenarios prominently features a long-tail distribution, where abundant (head) and rare (tail) objects coexist. The Open Long-Tail Recognition (OLTR) benchmark, as introduced in (Liu et al., 2019), tackles the dual challenges of imbalanced and few-shot learning within a single, streamlined training framework. Notably, robust recognition necessitates that head and tail classes share visual features to compensate for the sparse examples in tail classes, a concept underscored in (Liu et al., 2019; 2022). However, this sharing can inadvertently lead to confusion and erroneous predictions between visually similar objects also known as **inter-class bias** (Agarwal et al., 2022; Wang et al., 2019b), depicted in Figure 1. Moreover, the prevalence of head classes often biases models (He et al., 2016; Simonyan & Zisserman, 2015) towards them, adversely affecting performance on tail classes. The considerable variability within head classes, have been shown to generate local sub-centers (Deng et al., 2020) that intensify this bias and contribute to substantial **intra-class variance**, also showcased in Figure 1. To effectively navigate these complexities (further detailed in Section 2), *the underlying model must adeptly minimize both inter-class bias and intra-class variance*. We highlight in Section 2 that exist-

[1]Department of Computer Science, The University of Texas at Dallas, Richardson, TX, USA [2]Google Research, Sunnyvale, CA, USA [3]IBM Research, San Jose, CA, USA. Correspondence to: Anay Majee <anay.majee@utdallas.edu>, Rishabh Iyer <rishabh.iyer@utdallas.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]Project page: https://anaymajee.me/assets/project_pages/score.html.

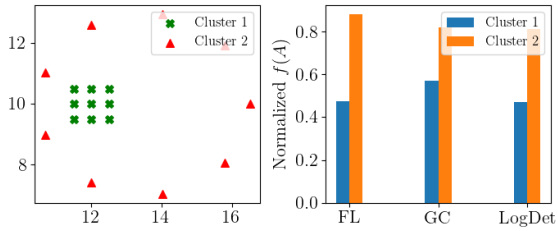


Figure 2. **Submodular functions** $f(A)$ in SCoRe model diversity (A = Cluster 1) and cooperation (A = Cluster 2).

ing approaches adopt metric/contrastive learners like SupCon (Khosla et al., 2020) to overcome the aforementioned challenges. Although State-of-the-Art (SoTA) approach SupCon shows appreciable variation to intra-class variance, as shown in Figure 1, it fails to apply a significant penalty to inter-class bias. This highlights the need for a joint objective that can penalize both pitfalls during model training.

To address these limitations, we introduce a **S**ubmodular **Co**mbinatorial **Re**presentation Learning (**SCoRe**) framework presenting a **family of combinatorial loss functions, as outlined in Table 1, designed to effectively address the dual challenges of intra-class variance and inter-class bias in long-tail recognition**. Our method leverages a combinatorial perspective by formulating the input dataset as a collection of sets (see Section 3), facilitating the use of set-based functions (Fujishige, 2005) as potent learning objectives. Specifically, SCoRe utilizes submodular functions, which model cooperation (Jegelka & Bilmes, 2011) (similarity) when minimized, and diversity (Lin & Bilmes, 2011; Kulesza, 2012) (dis-similarity) when maximized, due to their property of diminishing marginal returns (Fujishige, 2005). This is illustrated in Figure 2 where a submodular function $f(A)$ over a set A has a low value in the presence of low intra-class variance (cluster 1) modelling *cooperation* and has a high value otherwise (cluster 2) modeling *diversity*. We capitalize on these intrinsic properties of submodular functions to design a family of objectives as shown in Table 1 based on two well-known formulations- *Total Information* and *Total Correlation* which model total feature information and information gain on adding novel instances in a class respectively. Instances in SCoRe are driven by the strategic choice of submodular combinatorial functions as discussed in Section 3.1 where some objectives inherently demonstrate class-balancing (SCoRe-FL) while jointly modelling inter-class bias and intra-class variance. We demonstrate this in Figure 1(b, c) where SCoRe objectives exhibit larger relative variations in tackling both intra-class variance and inter-class bias compared to state-of-the-art approaches, making them superior for model training. The primary contributions of this paper are as follows:

- We introduce the novel SCoRe framework, introducing a **set-based combinatorial viewpoint** to representation learning under long-tail settings.

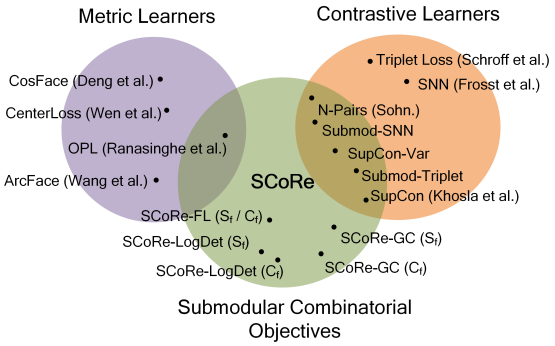


Figure 3. **Overview of Combinatorial Objectives** in SCoRe with respect to contrastive and metric learners.

- We show that **SCoRe generalizes several existing metric/contrastive learning approaches** like SupCon, N-pairs and OPL losses which are instances of SCoRe as shown in Figure 3 exhibiting combinatorial properties.

- Objectives in SCoRe **learn better features within fewer number of training epochs** in the *imbalanced* (Figure 6) setting while **demonstrating faster convergence** in the *balanced* setting as noted in Appendix A.4).

- Lastly, introduction of SCoRe objective functions result in outperforming State-of-the-Art (SoTA) approaches by **up to 7.6% for classification** tasks and **19.4% for object detection** tasks for class-imbalanced settings like CIFAR-10-LT, CIFAR-100-LT, MedMNIST, Imagenet-LT, IDD and LVIS (v1.0).

2. Related Work

Long-tail Learning: Visual recognition tasks in long-tail learning, which focus on learning from few over-represented and many under-represented classes, have traditionally tackled class imbalance in datasets through over-sampling of rare classes (Chawla et al., 2002) or under-sampling of abundant ones (Cui et al., 2019a; Zhang et al., 2021). However, these methods alter the original dataset's distribution, hindering model generalization. An alternative method is re-weighting class probabilities during learning, either by giving more weight to tail classes or restricting gradient updates for abundant classes (Shu et al., 2019; Wang et al., 2017; Zhou et al., 2020; Tan et al., 2020). Despite these efforts, recent research (Zhou et al., 2020) suggests that re-weighting strategies lead to sub-optimal representation learning. Progress has been made through metric/contrastive learning strategies (Cui et al., 2021; 2023; Li et al., 2021a; Zhu et al., 2022) and two-stage training objectives requiring extensive negative label information (Khosla et al., 2020; Chen et al., 2020a) which are challenging to implement on large datasets like ImageNet-LT (Liu et al., 2019). Vision transformers (Dosovitskiy et al., 2021), combined with

other techniques (Tian et al., 2022; Iscen et al., 2023), have shown significant improvements, albeit at a high computational cost. Recent SoTA approaches, such as (Cui et al., 2021; 2023; Du et al., 2023; 2024), blend data augmentation with contrastive learning. GPaCo innovates with parametric learnable class centers (Chen et al., 2020b), while GLMC (Du et al., 2023) introduces a loss combining global MixUp (Zhang et al., 2018), local CutMix (Yun et al., 2019), and a cumulative soft label reweighted loss. Notably, all SoTA methods employ some form of contrastive learning, underscoring its importance in long-tail learning.

Metric and Contrastive Learning: In supervised learning, traditional models using Cross-Entropy (CE) loss (Rumelhart et al., 1986) struggle with class imbalance and noisy labels. Metric learning approaches (Deng et al., 2019; Wang et al., 2018; Ranasinghe et al., 2021; Wang et al., 2019a) address this by learning distance (Schroff et al., 2015) or similarity (Deng et al., 2019; Wang et al., 2018) metrics, promoting orthogonality in feature space (Ranasinghe et al., 2021) and enhancing class-specific feature discrimination. Contrastive learning, derived from noise contrastive estimation (Gutmann & Hyvärinen, 2010), is prevalent in self-supervised learning (Chen et al., 2020a; He et al., 2020; Chen et al., 2020b) where label information is absent during training. In supervised domains, SupCon (Khosla et al., 2020) focuses on learning feature clusters, not just aligning features to centroids. Triplet loss (Schroff et al., 2015) contrasts one positive and negative pair, while N-pairs (Sohn, 2016) loss uses multiple negative pairs, and SupCon uses multiple positive and negative pairs. Lifted-Structure loss (Song et al., 2016) contrasts positives with the hardest negatives. SupCon is similar to Soft-Nearest Neighbors loss (Frosst et al., 2019), maximizing class entanglements. Despite successes, these methods rely on pairwise similarity metrics and may not ensure disjoint cluster formation.

Submodular Functions are set functions that satisfy a natural diminishing returns property. A set function $f : 2^{\mathcal{V}} \rightarrow \mathbb{R}$ (on a ground-set $\mathcal{V}$) is submodular if it satisfies $f(X) + f(Y) \geq f(X \cup Y) + f(X \cap Y), \forall X, Y \subseteq \mathcal{V}$ (Fujishige, 2005). These functions have been studied extensively in the context of data subset selection (Killamsetty et al., 2024; Kothawade et al., 2022b; Jain et al., 2023), active learning (Wei et al., 2015; Kothawade et al., 2022a; Beck et al., 2021; Kaushal et al., 2019b) and video-summarization (Kaushal et al., 2019a;c; 2021). Submodular functions are capable of modelling diversity, relevance, set-cover etc. which allows them to discriminate between different classes or slices of data while ensuring the preservation of most relevant features in each set. Very recent developments in the field have applied submodular functions like Facility-Location in metric learning (Oh Song et al., 2017). Minimizing submodular functions $f(A)$ over a set A models cooperation (Jegelka & Bilmes, 2011) between samples while maximizing submodular functions model diversity (Lin & Bilmes, 2011) making these functions suitable for learning diverse feature clusters in representation learning tasks which is yet to be studied in literature.

To the best of our knowledge , we are the first to demonstrate that combinatorial objectives using submodular functions are superior in creating tighter and well-separated feature clusters for representation learning. We are fore-runners in showing through Section 4 that SCoRe generalizes several existing contrastive learning approaches while some others can be reformulated as a superior submodular variant.

3. SCoRe: Submodular Combinatorial Representation Learning Framework

Supervised training for representation learning tasks proceed with learning a **feature extractor** $F(I, \theta)$ (Krizhevsky et al., 2012; Simonyan & Zisserman, 2015; He et al., 2016) which projects an input image I into a D dimensional feature space r, where $r = F(I, \theta) \in \mathbb{R}^D$, given parameters θ. A **classifier** $Clf(r, \theta)$ operates on the embeddings produced by F to categorize an input image I in the input dataset $\mathcal{T}$ into its corresponding class label c_i, where $i \in [1, 2, ...C]$. From our discussions in Section 1, representation learning in the long-tail setting requires the learning of generalizable features in $F(I, \theta)$, minimizing the impact of inter-class bias and intra-class variance. Previous literature in Section 2 has shown promise in this directions by promoting the learning of discriminative features by $F(I, \theta)$, attributed to be a function of a learning objective $L(\theta)$.

SCoRe introduces a family of **Combinatorial Loss Functions**, $L(\theta)$ as shown in Table 1, which trains the feature extractor F over all classes C in the dataset $\mathcal{T}$. SCoRe differs from existing approaches in the field through the introduction of a *set-based combinatorial viewpoint* by defining a dataset $\mathcal{T}$ as a collection of sets, $\mathcal{T} = \{A_1, A_2, \cdots, A_{|C|}\}$ over classes (now represented as sets A_k) in $\mathcal{T}$. Objectives in SCoRe further differ from previous works, by adopting submodular functions (Fujishige, 2005; Iyer et al., 2022) as learning objectives, which inherently model *cooperation* (Jegelka & Bilmes, 2011) and *diversity* (Lin & Bilmes, 2011) by the virtue of their diminishing marginal returns property (Fujishige, 2005). Adopting such a formulation in representation learning thus enforces cooperation within each class while promoting diversity between classes through $L(\theta)$, leading to the learning of discriminative class-specific features. Additionally, submodular functions **model the information contained in a set A (class in SCoRe) irrespective of its size**, motivating their application in long-tail settings. We adapt this viewpoint in Section 3.1 and propose two distinct formulations of combinatorial objectives - *Total Information*

Table 1. **Summary of various instantiations of SCoRe** and their respective combinatorial properties (detailed derivations in section A.7 of the appendix).

Objective Function	Equation $L(\theta)$	Combinatorial Property
Triplet Loss (Schroff et al., 2015)	$\sum_{k=1}^{\lvert C\rvert} \frac{1}{\lvert A_k\rvert}[\sum_{\substack{i,p\in A_k,\\ n\in\mathcal{V}\setminus A_k}} \max(0, D_{ip}^2(\theta) - D_{in}^2(\theta) + \epsilon)]$	Not Submodular
SNN (Frosst et al., 2019)	$\sum_{k=1}^{\lvert C\rvert} \frac{-1}{\lvert A_k\rvert} \sum_{i\in A_k} \log \sum_{j\in A_k} \exp(S_{ij}(\theta)) + \frac{1}{\lvert A_k\rvert} \log \sum_{\substack{i\in A_k\\ j\in\mathcal{V}\setminus A_k}} \exp(S_{ij}(\theta))$	Not Submodular
N-Pairs Loss (Sohn, 2016)	$\sum_{k=1}^{\lvert C\rvert} \frac{-1}{\lvert A_k\rvert} \sum_{i,j\in A_k} S_{ij}(\theta) + \frac{1}{\lvert A_k\rvert} \sum_{i\in A_k} log(\sum_{j\in\mathcal{V}} S_{ij}(\theta) - 1)$	Submodular
OPL (Ranasinghe et al., 2021)	$\sum_{k=1}^{\lvert C\rvert} \frac{1}{\lvert A_k\rvert}(1 - \sum_{i,j\in A_k} S_{ij}(\theta)) + \frac{1}{\lvert A_k\rvert} \sum_{i\in A_k} \sum_{j\in\mathcal{V}\setminus A_k} S_{ij}(\theta)$	Submodular
SupCon (Khosla et al., 2020)	$\sum_{k=1}^{\lvert C\rvert} [\frac{-1}{\lvert A_k\rvert} \sum_{i,j\in A_k} S_{ij}(\theta)] + \sum_{i\in A_k} \frac{1}{\lvert A_k\rvert} \log(\sum_{j\in V} exp(S_{ij}(\theta)) - 1)$	Submodular
Submod-Triplet	$\sum_{k=1}^{\lvert C\rvert} \frac{1}{\lvert A_k\rvert} \sum_{\substack{i\in A_k\\ n\in\mathcal{V}\setminus A_k}} S_{in}^2(\theta) - \sum_{i,p\in A} S_{ip}^2(\theta)$	Submodular
Submod-SNN	$\sum_{k=1}^{\lvert C\rvert} \frac{1}{\lvert A_k\rvert} \sum_{i\in A_k} [\log \sum_{j\in A_k} \exp(D_{ij}(\theta)) + \log \sum_{j\in\mathcal{V}\setminus A_k} \exp(S_{ij}(\theta))]$	Submodular
SupCon-Var	$\sum_{k=1}^{\lvert C\rvert} \frac{-1}{\lvert A_k\rvert} \sum_{i,j\in A_k} S_{ij}(\theta) + \frac{1}{\lvert A_k\rvert} \sum_{i\in A_k} log \sum_{j\in\mathcal{V}\setminus A_k} \exp(S_{ij}(\theta))$	Submodular
SCoRe-GC [S_f] (ours)	$\sum_{k=1}^{\lvert C\rvert} \frac{1}{\lvert A_k\rvert} [\sum_{i\in A_k} \sum_{j\in\mathcal{V}\setminus A_k} S_{ij}(\theta) - \lambda \sum_{i,j\in A_k} S_{ij}(\theta)]$	Submodular
SCoRe-GC [C_f] (ours)	$\sum_{k=1}^{\lvert C\rvert} \frac{\lambda}{\lvert A_k\rvert} \sum_{i\in A_k} \sum_{j\in\mathcal{V}\setminus A_k} S_{ij}(\theta)$	Submodular
SCoRe-LogDet [S_f] (ours)	$\sum_{k=1}^{\lvert C\rvert} \frac{1}{\lvert A_k\rvert} \log\det(S_{A_k}(\theta) + \lambda \mathbb{I}_{\lvert A_k\rvert})$	Submodular
SCoRe-LogDet [C_f] (ours)	$\sum_{k=1}^{\lvert C\rvert} \frac{1}{\lvert A_k\rvert} [\log\det(S_{A_k}(\theta) + \lambda \mathbb{I}_{\lvert A_k\rvert}) - \log\det(S_{\mathcal{V}}(\theta) + \lambda \mathbb{I}_{\lvert\mathcal{V}\rvert})]$	Submodular
SCoRe-FL [C_f/ S_f] (ours)	$\sum_{k=1}^{\lvert C\rvert} \frac{1}{\lvert\mathcal{V}\rvert} \sum_{i\in\mathcal{V}\setminus A_k} max_{j\in A_k} S_{ij}(\theta)$	Submodular

($L_{S_f}(\theta)$) and *Total Correlation* ($L_{C_f}(\theta)$), that inherently model inter-class bias and intra-class variance to learn compact yet well-separated class-specific feature clusters.

Training and evaluation of models through the SCoRe framework follows Khosla et al. (2020) and occurs in two stages. The first stage trains $F(I, \theta)$ to learn discriminative features through the newly introduced objectives $L(\theta)$, while the second stage trains the classifier $Clf(F(I, \theta))$ (F is frozen) using the standard cross-entropy loss (Rumelhart et al., 1986).

3.1. Combinatorial Loss Functions

Given an input data batch (referred to as ground set), $\mathcal{V} = \bigcup_{k=1}^{|C|} A_k$, and a submodular function $f(A_k; \theta)$ over a set A_k, we define a loss $L(\theta)$ which is an *instantiation of submodular information functions*. Here, A_k is a set representing each class in the dataset $\mathcal{T}$, $k \in [1, C]$ and f is defined with similarity kernels S, which depends on the parameters θ. We propose two flavors of information measures from (Iyer et al., 2022), the Total Submodular Information: $S_f(A_1, A_2, A_3, \ldots, A_{|C|}) = \sum_{k=1}^{|C|} f(A_k)$ and the Total Submodular Correlation: $C_f(A_1, A_2, A_3, \ldots, A_{|C|}) = \sum_{k=1}^{|C|} f(A_k) - f(\bigcup_{k=1}^{|C|} A_k)$. In context of representation learning, S_f captures the total information contained in an object class (referred to as set) $A_k \in \mathcal{T}$ while C_f captures the gain in information when new features are added to the set A_k. Using the aforementioned S_f and C_f formulations, we can define two variants of combinatorial loss functions $L(\theta)$ for long-tail recognition tasks:

$$
\begin{aligned}
L_{S_f}(\theta) &= \sum_{k=1}^{|C|} \frac{1}{N_f(A_k)} \; f(A_k; \theta), \\
L_{C_f}(\theta) &= \sum_{k=1}^{|C|} \frac{1}{N_f(A_k)} \left[f(A_k; \theta) - f(\bigcup_{k=1}^{|C|} A_k; \theta) \right]
\end{aligned}
\tag{1}
$$

Where, $N_f(A_k)$ is the normalization constant over each set in $\mathcal{T}$. As discussed in Section 3, minimizing a submodular function $f(A_k; \theta)$ through $L(\theta)$ over a set A_k captures *cooperation* (Jegelka & Bilmes, 2011; Iyer & Bilmes, 2015) between samples in the set, while maximizing $f(A_k; \theta)$ captures diversity / coverage (Lin & Bilmes, 2011; Iyer, 2015) between sets. Consequently, objective $L_{S_f}(\theta)$ **which minimizes S_f enforces intra-cluster compactness** by maximizing cooperation within each set A_k (minimizing $f(A_k; \theta)$ over each set A_k). Further, $L_{C_f}(\theta)$ **which minimizes C_f enforces both intra-cluster similarity (first term in L_{C_f}) and inter-cluster separation** (by maximizing $f(\cup_k A_k)$). L_{C_f} achieves inter-cluster separation by maximizing *diversity* between orthogonal sets in $\mathcal{T}$. We introduce several instantiations of the above formulations in our framework *based on the choice of the underlying submodular function* $f(A_k; \theta)$ producing a family of objective functions as shown in Table 1 for representation learning tasks.

3.1.1. Instantiations of Combinatorial Objectives in SCoRe

By varying the choice of submodular function $f(A)$ in SCoRe we propose several novel objective functions in Table 1. It is interesting to note that, our combinatorial objectives adopt a pairwise similarity kernel $S_{ij}(\theta)$ similar to existing approaches discussed in Section 2. However, SCoRe objectives utilize the similarity kernel only to compute *feature interactions between samples*, **differing from existing approaches in the aggregation of pairwise similarities to compute total information/ correlation** for a class $A_k \in \mathcal{T}$. In practice, we adopt the cosine similarity metric $S_{ij}(\theta)$ as used in Khosla et al. (2020), defined as $S_{ij}(\theta) = \frac{F(I_i,\theta)^{\mathrm{T}} \cdot F(I_j,\theta)}{||F(I_i,\theta)|| \cdot ||F(I_j,\theta)||}$ to produce three novel instantiations namely - SCoRe-FL, SCoRe-GC and SCoRe-LogDet based upon popular submodular functions in Iyer (2015); Kaushal et al. (2019a;c) - Graph-Cut, Facility-Location and Log-Determinant respectively.

SCoRe-FL based objective function minimizes the maximum similarity S_{ij} (where $i \neq j$) between orthogonal sets (different class labels).

Theorem 3.1. *If $f(A,\theta) = \sum_{i \in \mathcal{V}} \max_{j \in A} S_{ij}(\theta)$ represents the facility-location function over a set A then, $L_{S_f}(\theta)$ and $L_{C_f}(\theta)$ shown in Equation (2) represents the SCoRe-FL objective with $N_f(A_k) = |\mathcal{V}|$. Both $L_{S_f}(\theta)$ and $L_{C_f}(\theta)$ differ by a constant.*

$$
\begin{aligned}
L_{S_f}(\theta) &= \sum_{k=1}^{|C|} \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V} \setminus A_k} \max_{j \in A_k} S_{ij}(\theta) + 1, \\
L_{C_f}(\theta) &= \sum_{k=1}^{|C|} \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V} \setminus A_k} \max_{j \in A_k} S_{ij}(\theta)
\end{aligned}
\tag{2}
$$

Evident in its form in Equation (2), SCoRe-FL promotes large inter-cluster separation by minimizing the similarity between the hardest negative pair between $\mathcal{V} \setminus A_k$ and A_k.Additionally, **SCoRe-FL inherently introduces a class balancing** property making it an inevitable choice for learning in long-tail settings detailed in Section 3.1.3.

SCoRe-GC objective minimizes the feature similarity between representations of a positive set A_k and the remaining negative sets $\mathcal{V} \setminus A_k$ while maximizing the similarity among features in each set A_k.

Theorem 3.2. *If $f(A,\theta) = \sum_{i \in A, j \in V} S_{ij}(\theta) - \lambda \sum_{i,j \in A} S_{ij}(\theta)$ represents the Graph-Cut function over a set A then, $L_{S_f}(\theta)$ and $L_{C_f}(\theta)$ shown in Equation (3) represents the SCoRe-GC objective, normalized by $|A_k|$.*

$$
\begin{aligned}
L_{S_f}(\theta) &= \sum_{k=1}^{|C|} \frac{1}{|A_k|} \left[\sum_{\substack{i \in A_k, \\ j \in V \setminus A_k}} S_{ij}(\theta) - \lambda \sum_{i,j \in A_k} S_{ij}(\theta) \right], \\
L_{C_f}(\theta) &= \sum_{k=1}^{|C|} \frac{\lambda}{|A_k|} \sum_{\substack{i \in A_k, \\ j \in V \setminus A_k}} S_{ij}(\theta)
\end{aligned}
\tag{3}
$$

The hyper-parameter λ Lin & Bilmes (2011) controls the weightage of the loss to intra-class compactness over inter-cluster separation and is ablated upon in Table 7. It is interesting to note that **SCoRe-GC jointly models inter-cluster separation and intra-cluster compactness**. Minimizing the first term in SCoRe-GC ($-\sum_{i,j \in A_k} S_{ij}$) promotes intra-cluster compactness, while the second term ($\sum_{i \in A_k} \sum_{j \in \mathcal{V} \setminus A_k} S_{ij}(\theta)$) penalizes cluster overlaps i.e. promoting inter-class separation. Note that Orthogonal Projection Loss (OPL) and a version of Triplet Loss are special cases of of the GC based loss function.

SCoRe-LogDet presents an unique perspective by modelling the volume of a set A_k in the embedding space.

Theorem 3.3. *If $f(A,\theta) = \log\det(S_A(\theta) + \lambda \mathbb{I}_{|A|})$ represents the Log-Determinant function over a set A then, $L_{S_f}(\theta)$ and $L_{C_f}(\theta)$ shown in Equation (4) represents the SCoRe-LogDet objective which models the volume of each set $A_k \in \mathcal{T}$. $\mathbb{I}_{|A_k|}$ and $\mathbb{I}_{|\mathcal{V}|}$ indicate identity terms and $N_f(A_k) = |A_k|$, introduced for numerical stability.*

$$
\begin{aligned}
L_{S_f}(\theta) &= \sum_{k=1}^{|C|} \frac{1}{|A_k|} \log\det(S_{A_k}(\theta) + \lambda \mathbb{I}_{|A_k|}), \\
L_{C_f}(\theta) &= L_{S_f}(\theta) - \sum_{k=1}^{|C|} \frac{1}{|A_k|} \log\det(S_{\mathcal{V}}(\theta) + \lambda \mathbb{I}_{|\mathcal{V}|})
\end{aligned}
\tag{4}
$$

Minimizing SCoRe-LogDet over a set A_k through L_{S_f} **minimizes the cluster volume of A_k inherently reducing intra-class variance**. L_{C_f} (which empirically we see works better) captures both intra-cluster similarity and inter-cluster dissimilarity by additionally maximizing the diversity in the feature space $\mathcal{V}$. Note that the additive term $\mathbb{I}_{|A_k|}$ indicates an identity term introduced for numerical stability.

Proofs for all theorems are provided in Appendix A.6. Our experiments in Section 4 indicate that adopting set-based objectives defined in SCoRe outperforms existing metric/contrastive loss functions.

3.1.2. SCoRe Generalizes Existing Metric/Contrastive Learning Objectives

From the formulations in Table 1 we observe that **SCoRe generalizes to several metric/contrastive learning ob-**

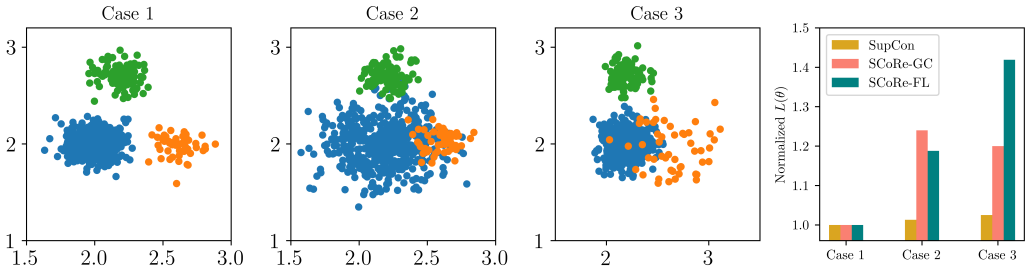


Figure 4. **Resilience to Intra-Class Variance and Inter-Class Bias under the Long-tail setting**. Case 1 demonstrates no intra-class variance and inter-class bias, while case 2 demonstrates larger variance for the head class wile case 3 demonstrates larger variance for the tail class inducing inter-cluster overlaps. The details of the experiment have been enclosed in Appendix A.2.

jectives which inherently demonstrate combinatorial properties. We provide proofs in Appendix A.7. N-pairs (Sohn, 2016) and OPL (Ranasinghe et al., 2021) and SupCon (Khosla et al., 2020) are all instances of SCoRe and are submodular (in the S_f form), exhibiting combinatorial properties. On the other hand, Triplet (Schroff et al., 2015) and SNN (Frosst et al., 2019) losses are not instances of SCoRe in their original form while their algebraically modified forms - Submod-Triplet and Submod-SNN as shown in Table 1 are instances of SCoRe and outperform their non-submodular counterparts.

Analyzing SoTA approach SupCon as shown in Table 1 we decompose it into an intra-class $(-\sum_{i,j\in A_k} S_{ij})$ and an inter-class $(\log(\sum_{i\in A_k, j\in V} exp(S_{ij}))$ term. The *inter-class term is computed over $\mathcal{V}$ which includes A_k, enforcing separation between samples in A_k which is counter intuitive* to the task at hand, hindering its capability to overcome inter-class bias. This is demonstrated by the low variation in the SupCon objective in the presence of large inter-class bias in Figure 1(b). To overcome this pitfall in SupCon we introduce, SupCon-Var (row 9 of Table 1) which minimizes the maximum similarity between a set A_k and $\mathcal{V} \setminus A_k$ in the inter-class term emerging as a stronger objective over SupCon in overcoming inter-class bias (the intra-class term is similar to SupCon).

3.1.3. Contrasting Instantiations of SCoRe

In this section we discuss some of the *properties* of combinatorial objectives, through experiments on synthetic data (refer Figure 4) which upholds our claim towards their application in long-tail recognition tasks. Although their properties largely depends on the choice of the submodular function $f(A_i; \theta)$ we enlist some unique ones favorable to long-tail settings.

- **SCoRe Objectives model Information in a Set**. Unlike exiting approaches, the novel formulations in SCoRe model total information (L_{S_f}) contained in a class (A_k) and the gain in information when new instances are added to A_k (L_{C_f}), *irrespective of the the size of the class*. This viewpoint in representation learning is novel to SCoRe.
- **SCoRe-FL demonstrates inherent Class-balancing**, equally weighting tail classes in contrast to head ones in model training. Unlike SoTA approaches (Khosla et al., 2020; Zhu et al., 2022; Chen et al., 2020a) that scale linearly with the size of the set $|A_k|$, SCoRe-FL has an inverse relation, and scales with respect to $\mathcal{V} \setminus A_k$ $(\sum_{i\in\mathcal{V}\setminus A_k} max_{j\in A_k} S_{ij}(\theta))$. This inherently introduces class balancing critical in long-tail settings. Figure 4 shows that, with increase in variance of the tail class (marked in orange) over the head class in case 3, SCoRe-FL and SCoRe-GC provide a larger relative penalty to the model than SoTA contrastive learners (Khosla et al., 2020; Chen et al., 2020a). Interestingly, the relative penalty applied by SCoRe-FL is significantly larger in case 3 than in case 2, where the variance of the head class (marked in blue) is larger than the tail. This highlights the class-balancing behavior in SCoRe-FL, thereby establishing its effectiveness in long-tail recognition tasks.
- **SCoRe-LogDet is a volumetric function** and models the volume of a feature cluster (Fujishige, 2005), minimizing which shrinks the cluster volume resulting in reduced intra-class variance (L_{S_f}). Additionally, maximizing the cluster volume over all sets (L_{C_f}) in the dataset maximizes diversity among clusters mitigating inter-class bias.

4. Experiments

We perform experiments on several long-tail vision benchmarks to show the effectiveness of objectives in SCoRe.

4.1. Datasets and Experimental Setup

CIFAR-10-LT consists of 10 disjoint classes with imbalance factors (IFs) ranging from 10 to 100 (Liu et al., 2019). The dataset is created by randomly sampling the balanced CIFAR-10 (Krizhevsky, 2009) dataset based on an exponentially decaying function with IF as rate of decay. Additionally, we introduce a pathological benchmark based on a *step* distribution by exploiting the hierarchy already available (living vs non-living objects) in the dataset. The data distributions of the adopted benchmarks are depicted in Figure 7(a, b).

Table 2. **Multi-class classification performance (Top1-Accuracy %) of combinatorial objectives in SCoRe** (shaded in Green) against existing approaches in Longtail recognition for *CIFAR-10-LT* and *CIFAR-100-LT* datasets for varying Imbalance Factors (IF).

	Method	CIFAR-10-LT			CIFAR-100-LT		
		IF=100	50	10	100	50	10
	CE	70.4	74.8	86.4	38.3	43.9	55.7
	Focal Loss (Lin et al., 2017)	70.38	76.72	86.66	38.41	44.32	55.78
Class/ Weight Balanced	BBN (Zhou et al., 2020)	79.82	82.18	88.32	42.56	47.02	59.12
	CB-Focal (Cui et al., 2019b)	74.6	79.3	87.1	39.6	45.2	58
	LogitAjust (Menon et al., 2021)	80.92	-	-	42.01	47.03	57.74
	weight balancing (Alshammari et al., 2022)	-	-	-	53.35	57.71	68.67
Augmentation Based	Mixup (Zhang et al., 2018)	73.06	77.82	87.1	39.54	54.99	58.02
	RISDA (Chen et al., 2022)	79.89	79.89	79.89	50.16	53.84	62.38
	CMO (Park et al., 2022)	-	-	-	47.2	51.7	58.4
Ensemble Classifier	RIDE (3 experts) + CMO (Wang et al., 2021)	-	-	-	50	53	60.2
	RIDE (3 experts) (Wang et al., 2021)	-	-	-	48.6	51.4	59.8
SSL-Pretraining	KCL (Kang et al., 2021)	77.6	81.7	88	42.8	46.3	57.6
	TSC (Li et al., 2022)	79.7	82.9	88.7	42.8	46.3	57.6
	SSD (Li et al., 2021b)	-	-	-	46.0	50.5	62.3
	BCL (Zhu et al., 2022)	84.32	87.24	91.12	51.93	56.59	64.87
	PaCo (Cui et al., 2021)	85.11	87.07	90.79	52.0	56.0	64.2
	PaCo + SCoRe-FL (ours)	**85.61**	**87.49**	**91.80**	**53.71**	**56.84**	**65.13**
One-Stage training	GLMC (Du et al., 2023)	88.50	91.04	94.90	58.0	63.78	73.43
	GLMC + SCoRe-GC (ours)	89.38	90.32	94.67	60.01	63.16	73.50
	GLMC + SCoRe-FL (ours)	**92.33**	**93.87**	**94.93**	**61.33**	**64.90**	**73.78**

CIFAR-100-LT is an extension of (Krizhevsky, 2009) with fine-grained labels. CIFAR-100-LT is also created by sampling its balanced counterpart based on an exponentially decaying function (IFs ranging from 10 to 100) but contains several few-shot classes (less than 20 instances per class).

ImageNet-LT introduced in (Liu et al., 2019), is a subset of the ImageNet (Deng et al., 2009) dataset consisting of 115.8K images from 1000 categories. The dataset shows severe imbalance following an exponentially decreasing distribution with a maximum and minimum of 1280 and 5 images per class.

MedMNIST (Yang et al., 2023) demonstrates real-world imbalance in medical datasets. We adopt the *OrganAMNIST* and *DermaMNIST* subsets of MedMNIST as they present extreme class-imbalance. OrganAMNIST consists of 41072 axial slices from CT volumes, highlighting 11 distinct organ structures while DermaMNIST contains 8010 samples of 7 different varieties of pigmented skin lesions. The data distributions of the adopted benchmarks are depicted in Figure 7(c, d).

IDD (Varma et al., 2019) depicts $\sim$ 41K real-world traffic scenarios characterized by longtail imbalance, high traffic density and large variability among object classes (Majee et al., 2021). IDD consists of $\sim$ 31K training examples and 10.2K validation images for the object detection task. Figure 8 depicts the data distribution of IDD highlighting the imbalance in the dataset.

LVIS (Gupta et al., 2019) dataset encapsulates 1203 commonplace objects from the MS-COCO (Lin et al., 2014) detection dataset (which consisted of 80 total classes) with extreme imbalance among classes. Every category in LVIS is assigned a distinct identifier from WordNet (Miller, 1995).

Table 3. **Longtail Recognition performance (Top-1 Acc %) on ImageNet-LT dataset**. We show that objectives in SCoRe(shaded in Green) generalize to existing SoTA approaches with improved overall performance. * indicates models trained with ResNet-50 as backbone.

Method	**ImageNet-LT**			
	Many	**Med**	**Few**	**All**
CE* (Baum & Wilczek, 1987)	64.0	33.8	5.8	41.6
SupCon* (Khosla et al., 2020)	53.4	2.9	0.0	22.0
CB-Focal* (Cui et al., 2019b)	39.6	32.7	16.8	33.2
LDAM* (Cao et al., 2019)	60.4	46.9	30.7	49.8
KCL* (Kang et al., 2021)	61.8	49.4	30.9	51.5
TSC* (Li et al., 2022)	63.5	49.7	30.4	52.4
BCL (Zhu et al., 2022)	67.9	54.2	**36.6**	57.1
RIDE (Wang et al., 2021) (3 experts)	66.4	53.9	35.6	56.2
PaCo (Cui et al., 2021) (400 epochs)	63.6	55.6	34.9	55.6
PaCo + SCoRe-GC (**ours**)	**69.4**	44.5	16.7	50.2
PaCo + SCoRe-FL (**ours**)	68.9	**55.8**	32.3	**57.5**
GLMC (Du et al., 2023)	70.1	52.4	30.4	56.1
GLMC + SCoRe-GC (**ours**)	68.7	52.4	31.5	55.7
GLMC + SCoRe-FL (**ours**)	**71.4**	**56.2**	**36.6**	**59.3**

The training dataset comprises a total of 100,000 images, encompassing 1.3 million instances, and the validation set contains 20,000 images.

For CIFAR-10-LT and CIFAR-100-LT experiments (refer Table 2) we follow the experimental setup of (Du et al., 2023) and adopt a ResNet-32 (He et al., 2016) backbone. We adopt the setup of PaCo (Cui et al., 2023) for the experiments on ImageNet-LT with a ResNeXt-50 (Xie et al., 2016) backbone. Additionally, for contrasting against existing metric/contrastive learners under long-tail settings of CIFAR-10 and MedMNIST (refer Table 5) we follow the experimental setup of (Khosla et al., 2020) with a ResNet-50 (He et al., 2016) backbone. Finally, we adopt the Faster-RCNN + FPN (Lin et al., 2017) architecture with a ResNet-101 back-

Table 4. **Object detection performance on IDD and LVIS datasets:** Applying our combinatorial objectives on a Faster-RCNN + FPN model produces the best Mean Average Precision (mAP) on real-world class-imbalanced settings.

Method	Backbone and head	mAP	mAP_{50}	mAP_{75}
India Driving Dataset (IDD)				
YOLO -V3[3](Redmon & Farhadi, 2018)	Darknet-53	11.7	26.7	8.9
Poly-YOLO[4] (Hurtík et al., 2020)	SE-Darknet-53	15.2	30.4	13.7
Mask-RCNN[4] (He et al., 2017)	ResNet-50	17.5	30.0	17.7
Retina-Net (Lin et al., 2017)	ResNet-50 + FPN	22.1	35.7	23.0
Faster-RCNN (Ren et al., 2015)	ResNet-101	27.7	45.4	28.2
Faster-RCNN + FPN	ResNet-101 + FPN	30.4	51.5	29.7
Faster-RCNN + SupCon	ResNet-101 + FPN	31.2	53.4	30.5
Faster-RCNN + SCoRe-GC [C_f]	ResNet-101 + FPN	33.6	56.0	34.6
Faster-RCNN + SCoRe-FL [S_f/C_f]	ResNet-101 + FPN	**36.3**	**59.5**	**37.1**
LVIS Dataset				
Faster-RCNN + FPN	ResNet-101 + FPN	14.2	24.4	14.9
Faster-RCNN + SupCon	ResNet-101 + FPN	14.4	26.3	14.3
Faster-RCNN + SCoRe-GC [C_f]	ResNet-101 + FPN	17.7	29.1	18.3
Faster-RCNN + SCoRe-FL [S_f/C_f]	ResNet-101 + FPN	**19.1**	**30.5**	**20.3**

bone in the Detectron2[2] framework for experiments on IDD and LVIS datasets. We train all our models on 2 NVIDIA A6000 GPUs with code released at `https://github.com/amajee11us/SCoRe.git`. More details on the datasets, experimental setup and hyper-parameters for individual experiments in section A.2 of the appendix.

4.2. Results on Long-tail Image Classification

Benchmark Results: At first, we conduct experiments on the long-tail benchmarks of CIFAR-10-LT, CIFAR-100-LT and ImageNet-LT outlined in (Liu et al., 2019). To the SoTA approaches in long-tail recognition - PaCo and GLMC, we introduce the proposed combinatorial objectives in SCoRe as auxiliary supervision head. For the results on **CIFAR-10-LT** and **CIFAR-100-LT** we adopt the benchmark published in GLMC (Du et al., 2023) with addition of combinatorial objectives as tabulated in Table 2. We compare objectives in SCoRe against class balancing (row 2,3) / weight balancing techniques (row 4,5), augmentation based approaches (rows 6, 7, 8) and approaches adopting auxiliary supervision (mostly self-supervised) in training (rows 9 - 14). We observe that addition of SCoRe objectives, specifically SCoRe-FL improves overall performance across various degrees of imbalance (denoted as IF in Table 2), upto 1.11% over PaCo and 4.33% over GLMC (IF=100) for the CIFAR-10-LT dataset. In particular, the gain in performance is significant under severe imbalance (IF=100) whereas, in a more balanced setting (IF=10) the performance boost is incremental. We show that the aforementioned observation continues to hold in CIFAR-100-LT (containing fine-grained labels and few-shot classes) with 3.29% gain over PaCo and 5.74% gain over GLMC (IF=100), where introduction of combinatorial objectives boosts performance in all settings significantly on ones with large imbalance (IF=100).

Secondly, experiments conducted on **ImageNet-LT** shown in Table 3 clearly demonstrates the supremacy of objectives

[2]`https://github.com/facebookresearch/detectron2`

[3]Results are from (Hurtík et al., 2020).

Table 5. **Multi-class classification performance (Top1-Accuracy %) of combinatorial objectives in SCoRe** (shaded in Green) against existing approaches in metric learning and their submodular instances (shaded in blue) on **Class-Imbalanced** *CIFAR-10-LT* (columns 2 - 3) and *MedMNIST* (columns 4 - 5) datasets.

Objective Function	CIFAR-10-LT		MedMNIST	
	LongTail IF=10	Pathological Step	OrganMNIST (Axial)	Derma MNIST
Cross-Entropy (CE)	86.44	74.49	81.80	71.32
Triplet Loss (Schroff et al., 2015)	85.94	74.23	81.10	70.92
N-Pairs (Sohn, 2016)	89.70	73.10	84.84	71.82
Lifted Structure Loss (Song et al., 2016)	82.86	73.98	84.55	71.62
SNN (Frosst et al., 2019)	83.65	75.97	83.85	71.87
Multi-Similarity Loss (Wang et al., 2019a)	82.40	76.72	85.50	71.02
SupCon (Khosla et al., 2020)	89.96	78.10	87.35	72.12
Submod-Triplet (ours)	89.20	74.36	86.03	72.35
Submod-SNN (ours)	89.28	78.76	86.21	71.77
SupCon-Var (ours)	90.81	81.31	87.48	72.51
SCoRe-GC [S_f] (ours)	89.20	76.89	86.28	69.10
SCoRe-GC [C_f] (ours)	90.83	87.37	**87.57**	72.82
SCoRe-LogDet [C_f] (ours)	90.80	87.00	87.00	72.04
SCoRe-FL [C_f/ S_f] (ours)	**91.80**	**87.49**	87.22	**73.77**

in SCoRe which demonstrate combinatorial properties by outperforming SoTA approaches like PaCo (Cui et al., 2023) and GLMC (Du et al., 2023) by 2.1% (*All* category) and 3.6% respectively. Introduction of combinatorial objectives like SCoRe-FL *significantly improves performance on the tail classes without significant loss in performance on head ones* reinstating the inherent *class-balancing* property of SCoRe-FL. However, we observe a drop in performance in performance in the few-shot (*Few*) and comparable performance in many shot (*Many*) settings. This can be attributed to the objectives guiding the model to overfit on the tail classes during training. For all datasets in the aforementioned benchmarks, it is interesting to note that SCoRe objectives augment the underlying architecture in existing SoTA approaches (Cui et al., 2023) thus *demonstrating the generalizability of our objectives* in this domain.

Generalization to Existing Metric/Contrastive Learners: Table 5 compares the performance of the newly introduced objectives in SCoRe against existing metric/contrastive learners on the framework in (Khosla et al., 2020). We conduct experiments on the *Longtail* (IF=10) and *step* distributions of the **CIFAR-10** dataset (refer Appendix A.2). The combinatorial objectives in SCoRe, namely the SCoRe-FL objective shows a 2% and 7.6% improvement over SupCon for the *longtail* and *step* distributions respectively. Even the reformulated submodular objectives - *Submod-Triplet*, *Submod-SNN* and *SupCon-Var* demonstrate upto 3.5%, 3.7% and 4.11% respectively over their non-submodular counterparts. Further, we demonstrate the effectiveness of the SCoRe objectives on the naturally imbalanced *OrganAMNIST* and *DermaMNIST* subsets of **MedMNIST** in Table 5 (columns 4,5). SCoRe objectives outperform SoTA approaches by 0.25% (as shown by SCoRe-GC over SupCon) and 1.5% (as shown by SCoRe-FL over SupCon) for OrganMNIST and DermaMNIST respectively. Similar to CIFAR-10 benchmark we also observe that, submodular counterparts of existing contrastive losses consistently outperform their non-submodular counterparts.

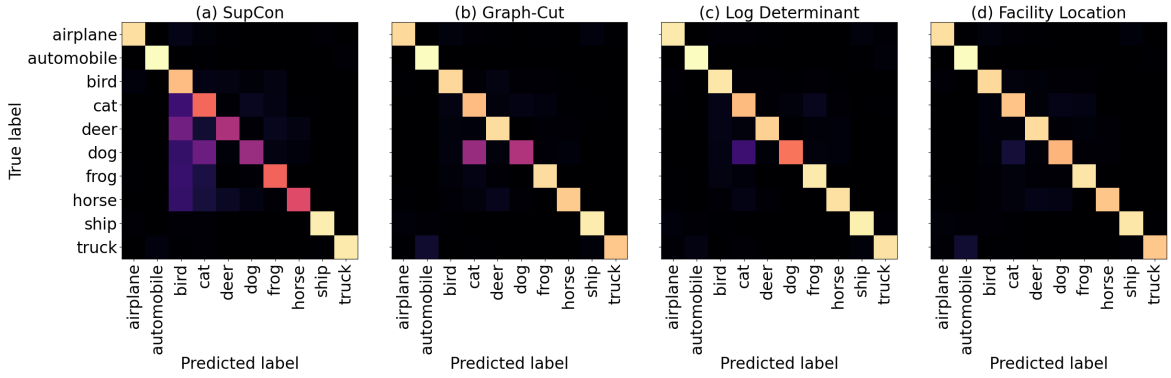


Figure 5. **Comparison of Confusion Matrix plots** between (a) SupCon (Khosla et al., 2020), (b) Graph-Cut (GC), (c) Log Determinant, and (d) Facility Location (FL) for the longtail imbalanced setting of CIFAR-10 dataset. We show a significant reduction in inter-class bias when employing combinatorial objectives in SCoRe characterized by reduced confusion between classes.

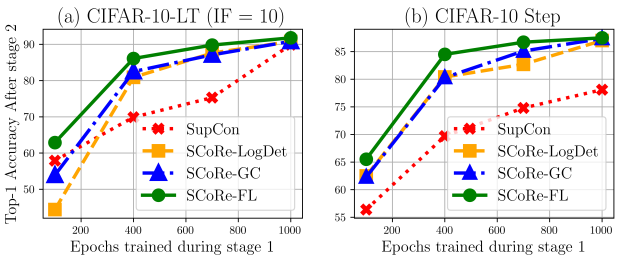


Figure 6. **SCoRe demonstrates faster learning of discriminative representations** in CIFAR-10 with fewer training epochs for both *longtail* and *step* imbalanced settings.

Discriminative Clustering and Convergence: We compare confusion matrix plots on predicted class labels after stage 2 of model training for the CIFAR-10-LT (IF=10) dataset. Plots in Figure 5 show that SupCon shows $\sim 22\%$ overall confusion with elevated confusion between the *animal* hierarchy of CIFAR-10. Both SCoRe-GC and SCoRe-LogDet demonstrate confusion between structurally similar objects like *cat* and *dog* (4-legged animals). Interestingly, a significant drop in confusion is observed when adopting SCoRe-FL with a minimum of 8.2%. The reduction in confusion by objectives proposed in SCoRe **points to a reduction in inter-class bias** (Majee et al., 2021). This is correlated to reducing the impact of class-imbalance due to formation of discriminative feature clusters.

Finally, from Figure 6 we show that SCoRe facilitates the learning **robust feature representations within significantly fewer training epochs**. Although we see a larger performance gain in the long-tail setting (SCoRe-FL) (Figure 6(a,b)), SCoRe demonstrates a *faster convergence in the balanced setting* as well as discussed in Appendix A.4.

4.3. Results on Class-Imbalanced Object Detection

We benchmark the performance of our approach against SoTA object detectors which adopt Focal Loss (Lin et al., 2017), data-augmentations etc. At first, we introduce a contrastive learning based objective (SupCon) in the box classification head of the object detector and show that contrastive learning outperforms standard model training (using CE loss) on IDD by 3.6 % (1.9 mAP_{50} points) and 10.6% (2.8 mAP_{50} points) on LVIS datasets. Secondly, we introduce the objectives in SCoRe to the box classification head and show that they outperform the SoTA as well as the contrastive learning objective (Khosla et al., 2020). The results in Table 4 show that the SCoRe-FL and SCoRe-GC objectives outperform the SoTA method by 6.1 mAP_{50} and 2.6 mAP_{50} points respectively on IDD, alongside 6.1 mAP_{50} and 4.7 mAP_{50} points respectively on LVIS. Additionally, from the class-wise performance on IDD as shown in Figure 8, SCoRe objectives demonstrate a sharp rise in performance (mAP_{50} value) of the rare classes (a maximum of 6.4 mAP points for *Bicycle* class) over SoTA objectives.

5. Conclusion

The SCoRe framework introduces a novel family of combinatorial objectives based on submodular information measures, designed to address long-tail imbalance in real-world vision tasks. A key strength of SCoRe is its ability to present new methodologies alongside generalizing to existing metric and contrastive learners. Empirically, SCoRe demonstrates remarkable effectiveness, achieving up to 5.74% improvement in classification tasks on datasets like CIFAR-10-LT, CIFAR-100-LT and MedMNIST, 2.1% on ImageNet-LT and an impressive 19.4% enhancement in object detection on challenging datasets such as the Indian Driving Dataset (IDD) and LVIS. The integration of combinatorial counterparts of existing objectives further underscores SCoRe's versatility, leading to significant performance gains and validating its efficacy in managing class imbalance. SCoRe's bridging of novel and established learning strategies marks a substantial contribution to the field, offering a robust solution for real-world applications.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning in general and Representation Learning in particular. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

We gratefully thank anonymous reviewers for their valuable comments. We would also like to extend our gratitude to our fellow researchers from the CARAML lab at UT Dallas - Nathan Beck and Truong Pham for their suggestions. This work is supported by the National Science Foundation under Grant Numbers IIS-2106937, a gift from Google Research, an Amazon Research Award, and the Adobe Data Science Research award. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, Google or Adobe.

References

Agarwal, A., Majee, A., Subramanian, A., and Arora, C. Attention guided cosine margin to overcome class-imbalance in few-shot road object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pp. 221–230, 2022. doi: 10.1109/WACVW54805.2022.00028.

Alshammari, S., Wang, Y., Ramanan, D., and Kong, S. Long-tailed recognition via weight balancing. In *CVPR*, 2022.

Baum, E. and Wilczek, F. Supervised learning of probability distributions by neural networks. In *Neural Information Processing Systems*, 1987.

Beck, N., Sivasubramanian, D., Dani, A., Ramakrishnan, G., and Iyer, R. K. Effective evaluation of deep active learning on image classification tasks. *ArXiv*, abs/2106.15324, 2021.

Cao, K., Wei, C., Gaidon, A., Aréchiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing, NeurIPS*, 2019.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: Synthetic minority over-sampling technique. 16(1):321–357, 2002.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. *Intl. Conf. on Machine Learning (ICML)*, 2020a.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Chen, X., Zhou, Y., Wu, D., Zhang, W., Zhou, Y., Li, B., and Wang, W. Imagine by reasoning: A reasoning-based implicit semantic data augmentation for long-tailed classification. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

Cui, J., Zhong, Z., Liu, S., Yu, B., and Jia, J. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 715–724, 2021.

Cui, J., Zhong, Z., Tian, Z., Liu, S., Yu, B., and Jia, J. Generalized parametric contrastive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–12, 2023.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019a.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. Class-balanced loss based on effective number of samples. In *CVPR*, 2019b.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

Deng, J., Guo, J., Liu, T., Gong, M., and Zafeiriou, S. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *European Conference on Computer Vision*, 2020.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Du, C., Wang, Y., Song, S., and Huang, G. Probabilistic contrastive learning for long-tailed visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Du, F., Yang, P., Jia, Q., Nan, F., Chen, X., and Yang, Y. Global and local mixture consistency cumulative learning

for long-tailed visual recognitions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15814–15823, June 2023.

Frosst, N., Papernot, N., and Hinton, G. E. Analyzing and improving representations with the soft nearest neighbor loss. In *International Conference on Machine Learning*, 2019.

Fujishige, S. *Submodular Functions and Optimization*, volume 58. Elsevier, 2005.

Gupta, A., Dollar, P., and Girshick, R. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for Image Recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask R-CNN. In *IEEE Intl. Conf. on Computer Vision (ICCV)*, 2017.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Hurtík, P., Molek, V., Hula, J., Vajgl, M., Vlasánek, P., and Nejezchleba, T. Poly-YOLO: Higher speed, more precise detection and instance segmentation for yolov3. *ArXiv*, abs/2005.13243, 2020.

Iscen, A., Fathi, A., and Schmid, C. Improving image recognition by retrieving from web-scale image-text data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19295–19304, June 2023.

Iyer, R. and Bilmes, J. Polyhedral aspects of submodularity, convexity and concavity. *arXiv preprint arXiv:1506.07329*, 2015.

Iyer, R., Khargonkar, N., Bilmes, J., and Asnani, H. Generalized submodular information measures: Theoretical properties, examples, optimization algorithms, and applications. *IEEE Transactions on Information Theory*, 68 (2):752–781, 2022.

Iyer, R. K. *Submodular optimization and machine learning: Theoretical results, unifying and scalable algorithms, and applications*. PhD thesis, 2015.

Jain, E., Nandy, T., Aggarwal, G., Tendulkar, A. V., Iyer, R. K., and De, A. Efficient data subset selection to generalize training across models: Transductive and inductive networks. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Jegelka, S. and Bilmes, J. Submodularity beyond submodular energies: Coupling edges in graph cuts. In *CVPR 2011*, 2011.

Kang, B., Li, Y., Xie, S., Yuan, Z., and Feng, J. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2021.

Kaushal, V., Iyer, R., Doctor, K., Sahoo, A., Dubal, P., Kothawade, S., Mahadev, R., Dargan, K., and Ramakrishnan, G. Demystifying multi-faceted video summarization: Tradeoff between diversity, representation, coverage and importance. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 452–461, 2019a.

Kaushal, V., Iyer, R., Kothawade, S., Mahadev, R., Doctor, K., and Ramakrishnan, G. Learning from less data: A unified data subset selection and active learning framework for computer vision. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019b.

Kaushal, V., Subramanian, S., Kothawade, S., Iyer, R., and Ramakrishnan, G. A framework towards domain specific video summarization. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pp. 666–675. IEEE, 2019c.

Kaushal, V., Kothawade, S., Tomar, A., Iyer, R., and Ramakrishnan, G. How good is a video summary? a new benchmarking dataset and evaluation framework towards realistic video summarization, 2021.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, 2020.

Killamsetty, K., S, D., Ramakrishnan, G., De, A., and Iyer, R. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, 2021.

Killamsetty, K., Abhishek, G. S., Aakriti, Ramakrishnan, G., Evfimievski, A. V., Popa, L., and Iyer, R. AUTOMATA: gradient based data subset selection for compute-efficient

hyper-parameter tuning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2024.

Kothawade, S., Ghosh, S., Shekhar, S., Xiang, Y., and Iyer, R. K. Talisman: Targeted active learning for object detection with rare classes and slices using submodular mutual information. In *Computer Vision - ECCV 2022 - 17th European Conference*, 2022a.

Kothawade, S., Kaushal, V., Ramakrishnan, G., Bilmes, J. A., and Iyer, R. K. PRISM: A rich class of parameterized submodular information measures for guided data subset selection. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI*, pp. 10238–10246, 2022b.

Krizhevsky, A. Learning multiple layers of features from tiny images. 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

Kulesza, A. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012. ISSN 1935-8245.

Li, T., Wang, L., and Wu, G. Self supervision to distillation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 630–639, 2021a.

Li, T., Wang, L., and Wu, G. Self supervision to distillation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 630–639, 2021b.

Li, T., Cao, P., Yuan, Y., Fan, L., Yang, Y., Feris, R., Indyk, P., and Katabi, D. Targeted supervised contrastive learning for long-tailed recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Lin, H. and Bilmes, J. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

Lin, T., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal Loss for dense object detection. In *IEEE Intl. Conf. on Computer Vision (ICCV)*, pp. 2999–3007, 2017.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common Objects In Context. In *ECCV*, pp. 740–755, 2014.

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., and Yu, S. X. Large-scale long-tailed recognition in an open world. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., and Yu, S. X. Open long-tailed recognition in a dynamic world. *TPAMI*, 2022.

Majee, A., Agrawal, K., and Subramanian, A. Few-Shot Learning For Road Object Detection. In *AAAI Workshop on Meta-Learning and MetaDL Challenge*, volume 140, pp. 115–126, 2021.

Menon, A. K., Jayasumana, S., Rawat, A. S., Jain, H., Veit, A., and Kumar, S. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=37nvvqkCo5.

Miller, G. A. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

Oh Song, H., Jegelka, S., Rathod, V., and Murphy, K. Deep metric learning via facility location. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Park, S., Hong, Y., Heo, B., Yun, S., and Choi, J. Y. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2022.

Ranasinghe, K., Naseer, M., Hayat, M., Khan, S., and Khan, F. S. Orthogonal projection loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Redmon, J. and Farhadi, A. YOLOv3: An Incremental Improvement. *CoRR*, abs/1804.02767, 2018. URL http://arxiv.org/abs/1804.02767.

Ren, S., He, K., Girshick, R. B., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., and Meng, D. Meta-weight-net: Learning an explicit mapping for sample weighting. In *NeurIPS*, 2019.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *Intl. Conf. on Learning Representations*, 2015.

Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Inf. Processing Systems*, 2016.

Song, H. O., Xiang, Y., Jegelka, S., and Savarese, S. Deep metric learning via lifted structured feature embedding. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

Sun, B., Li, B., Cai, S., Yuan, Y., and Zhang, C. Fsce: Few-shot object detection via contrastive proposal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2021.

Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., and Yan, J. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Tian, C., Wang, W., Zhu, X., Wang, X., Dai, J., and Qiao, Y. Vl-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition. In *Proceedings of the European Conference in Computer Vision (ECCV)*, 2022.

Varma, G., Subramanian, A., Namboodiri, A., Chandraker, M., and Jawahar, C. V. IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pp. 1743–1751, 2019. doi: 10.1109/WACV.2019.00190.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., and Liu, W. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Wang, X., Han, X., Huang, W., Dong, D., and Scott, M. R. Multi-similarity loss with general pair weighting for deep metric learning. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2019a.

Wang, X., Lian, L., Miao, Z., Liu, Z., and Yu, S. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021.

Wang, Y.-X., Ramanan, D., and Hebert, M. Learning to model the tail. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Wang, Y.-X., Ramanan, D., and Hebert, M. Meta-Learning To Detect Rare Objects. In *ICCV*, 2019b.

Wei, K., Iyer, R., and Bilmes, J. Submodularity in data subset selection and active learning. In *ICML*, 2015.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016.

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., and Ni, B. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Zhang, S., Li, Z., Yan, S., He, X., and Sun, J. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2361–2370, June 2021.

Zhou, B., Cui, Q., Wei, X.-S., and Chen, Z.-M. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. pp. 1–8, 2020.

Zhu, J., Wang, Z., Chen, J., Chen, Y.-P. P., and Jiang, Y.-G. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6908–6917, 2022.

A. Appendix

A.1. Notations

Following the problem definition in Section 3 we introduce several important notations in Table 6 that are used throughout the paper.

Table 6. Collection of notations used in the paper.

Symbol	Description
$\mathcal{T}$	The training Set. $\lvert\mathcal{T}\rvert$ denotes the size of the training set.
$\mathcal{V}$	Ground set containing feature vectors from all classes in $\mathcal{T}$.
$F(x, \theta)$	Convolutional Neural Network used as feature extractor.
$Clf(.,.)$	Multi-Layer Perceptron as classifier.
θ	Parameters of the feature extractor.
$S_{ij}(\theta)$	Similarity between images $i, j \in \mathcal{T}$.
$D_{ij}(\theta)$	Distance between images $i, j \in \mathcal{T}$.
p	Positive sample which is of the same class c_i as the anchor a.
n	Negative sample which is of the same class c_i as the anchor x.
A_k	Target set containing feature representation from a single class $k \in c_i$.
$f(A)$	Submodular Information function over a set A.
S_f	Variant of submodular information function denoting total information in the ground set V.
C_f	Variant of submodular information function denoting total correlation in the ground set V.
$L(\theta)$	Loss value computed over all classes $c_i \in C$.
$f(A_k, \theta)$	Instantiation of objective functions in SCoRe over a set/class A_k given parameters θ.
AK	Actinic Keratoses
BCC	Basal Cell Carcinoma
KLL	Keratosis-Like-Lesions
DF	Dermatofibroma
M	Melanoma
MN	Melanocytic Nevi
VL	Vascular Lesions

A.2. Experimental Setup : Additional Information

In this section we iron out the dataset details, training and inference settings of various datasets/tasks encompassed in the SCoRe framework.

A.2.1. Class-Imbalanced Image Classification on Cannonical benchmarks

CIFAR-10-LT and CIFAR-100-LT: Following the discussion on the choices of datasets introduced by the OLTR (Liu et al., 2019) benchmark in Section 4 we vary the Imbalance Factors (IF) of an exponentially decaying function to sample the CIFAR-10 and CIFAR-100 datasets to create their respective long-tail counterparts. We vary the imbalance factors between 10, 50 and 100 to produce pathologically imbalanced datasets. The higher the value of IF, the larger is the number of tail classes a sample of which for IF=10 is depicted in Figure 7. Additionally we introduce a **step** function based imbalance setting which exploits the hierarchy already available in CIFAR-10. The CIFAR-10 dataset can be broadly classified into *animal* and *automobile* classes. We use this information to subsample the *animal* (chosen at random) class objects to create an imbalanced step data distribution. The distributions of the dataset is depicted in Figure 7.

For contrasting against SoTA metric/contrastive learners in Table 5, we train our models by adopting the training strategy of (Khosla et al., 2020) and release the codebase at `https://github.com/amajee11us/SCoRe.git`. For stage 1 we train a ResNet-50 backbone with a batch size of 512 (1024 after augmentations) with an initial learning rate of 0.4, trained for 1000 epochs with a cosine annealing scheduler and a temperature for the combinatorial objectives to be 0.7. In stage 2 we freeze the backbone and use the output of the final pooling layer to train a linear classifier Clf with a batch size of 512 and a constant learning rate of 0.8.

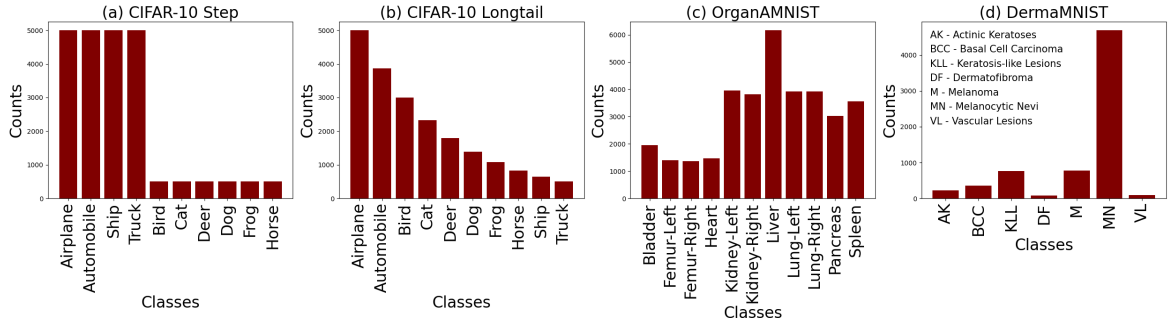


Figure 7. **Data Distribution of CIFAR-10 and MedMNIST datasets** under the class-imbalanced setting. Subfigure (a,b) depicts the longtail and step based pathological class imbalanced settings in CIFAR-10 and (c,d) depicts the naturally imbalanced OrganAMNIST and DermaMNIST[6]datasets respectively.

Further, for contrasting against existing SoTA approaches in Long-Tail recognition as shown in Table 2 for both CIFAR-10 and CIFAR-100 datasets, we introduce the objectives defined in SCoRe into the existing frameworks of SoTA approaches like GLMC (Du et al., 2023) and PaCo (Cui et al., 2021). For PaCo we replace its contrastive objective - MoCo (He et al., 2020) with the combinatorial objectives in SCoRe, while for GLMC we introduce our novel objectives to contrast between local and global features in the original architecture.

ImageNet-LT : Unlike CIFAR-10, ImageNet (Deng et al., 2009) dataset presents a large scale image recognition benchmark. Similar to CIFAR-10 authors in (Liu et al., 2019) subsample this dataset to introduce a pathological imbalance. In contrast to CIFAR-10 ImageNet contains 1000 classes and the longtail version ImageNet-LT contains severe imbalance with a maximum of 1280 and minimum of 5 instances for a particular class in the dataset[4].

For the ImageNet-LT dataset we follow the training and inference strategy adopted by PaCo (Cui et al., 2023) with modifications to the objective functions which are released at `https://github.com/amajee11us/SCoRe/blob/main/objectives/combinatorial/PaCoFL.py` (for SCoRe-FL loss function). Unlike the CIFAR-10 dataset PaCo adopts a one stage training strategy where we train the model for 400 epochs on a ResNeXt-50 (Xie et al., 2016) backbone with an initial learning rate of 0.1 and a cosine annealing scheduler. The model is trained on 4 GPUs, with a batch size of 32 on each GPU and a temperature for the combinatorial objectives to be 0.7. The results on the longtail distribution of ImageNet-LT benchmarked against several approaches in Longtail learning have been depicted in Table 3.

A.2.2. Class-Imbalanced Medical Image Classification

In contrast to pathological imbalance introduced in cannonical benchmarks we conduct our experiments on two subsets of **MedMNIST** (Yang et al., 2023) dataset which demonstrate a natural class-imbalanced setting.

OrganAMNIST dataset consists of axial slices from CT volumes, highlighting 11 distinct organ structures for a multi-class classification task. Each image is of size $[1 \times 28 \times 28]$ pixels. The OrganAMNIST dataset contains 34581 training and 6491 validation samples of single channel images highlighting various modalities of 8 different organs.

DermaMNIST subset presents dermatoscopic images of pigmented skin lesions, also resized to $[3 \times 28 \times 28]$ pixels. This dataset supports a multi-class classification task with 7 different dermatological conditions. Although the DermaMNIST has RGB images, it is a small scale dataset with a total of 7007 training samples and 1003 validation samples.

Similar to CIFAR-10 we train our models on MedMNIST datasets using the two stage training strategy outlined in SCoRe at `https://github.com/amajee11us/SCoRe.git`. For both these subsets used in our framework, pixel values were normalized to the range $[0, 1]$, and we relied on the standard train-test splits provided with the datasets for our evaluations. The results from the experiments are discussed in Section 4. For both data subsets we train the model for 500 epochs in stage-1 with a ResNet-50 backbone, a batch size of 256 and a cosine annealing scheduler. For stage-2 we use the frozen feature extractor and train a classifier with a batch size of 128 for 100 epochs with early stopping.

[4]The dataset as been adapted from `https://liuziwei7.github.io/projects/LongTail.html`.
[6]Abbreviations are included in section A.1 of the appendix.

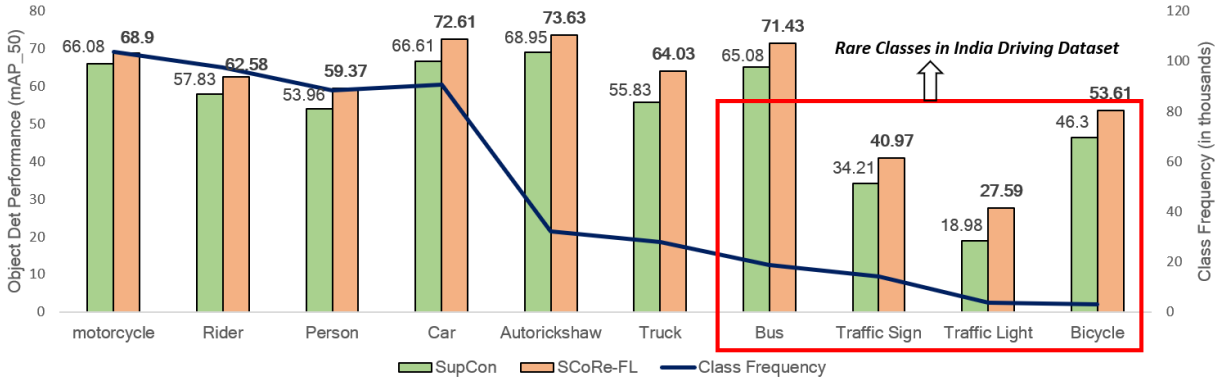


Figure 8. **The effect of class-imbalance** on the performance metrics (mAP_{50}) for the object detection task of the India Driving Dataset (IDD). As the class frequency (shown as blue line) decreases, proposed objectives in SCoRe (shown in red) consistently outperforms SoTA approaches like SupCon (shown in green) in detecting rare road objects like *bicycle*, *traffic light* etc. in IDD.

A.2.3. Class-Imbalanced Object Detection

IDD-Detection (Varma et al., 2019) dataset demonstrates an unconstrained driving environment, characterized by natural class-imbalance, high traffic density and large variability among object classes. This results in the presence of rare classes like *autorickshaw*, *bicycle* etc. and small sized objects like *traffic light*, *traffic sign* etc. There are a total of 31k training images in IDD and 10k validation images of size $[3 \times 1920 \times 1080]$ with high traffic density, occlusions and varying road conditions.

The architecture of the object detector is a Faster-RCNN (Ren et al., 2015) model with a ResNet-101 backbone alongside the Feature Pyramidal Network (FPN) as in (Lin et al., 2017) to handle varying object sizes. Our framework also draws inspiration from FSCE (Sun et al., 2021) with proposed modifications to the objective functions. We initialize our Faster-RCNN + FPN model with pretrained weights from a ImageNet trained model and fine-tune the model on IDD / LVIS (full dataset). We keep the Region Proposal Network (RPN) and the ROI pooling layers unfrozen to adapt to the rare classes. We also double the maximum number of proposals kept after Non-Maximal Suppression (NMS), bringing more proposals from rare classes to the foreground. We consider only half the number of proposals from the ROI pooling layer (top 256 out of 512) for computing the loss function. This forces the objective function to better penalize the object detector for predicting low objectness scores for objects belonging to the rare classes. The model is trained for 17000 iterations with a batch size of 8 and an initial learning rate of 0.02. A step based learning rate scheduler is adopted to reduce the learning rate by 10x at 12000 and 15000 iterations respectively. The results for the class-wise performance on the IDD dataset is depicted in Figure 8.

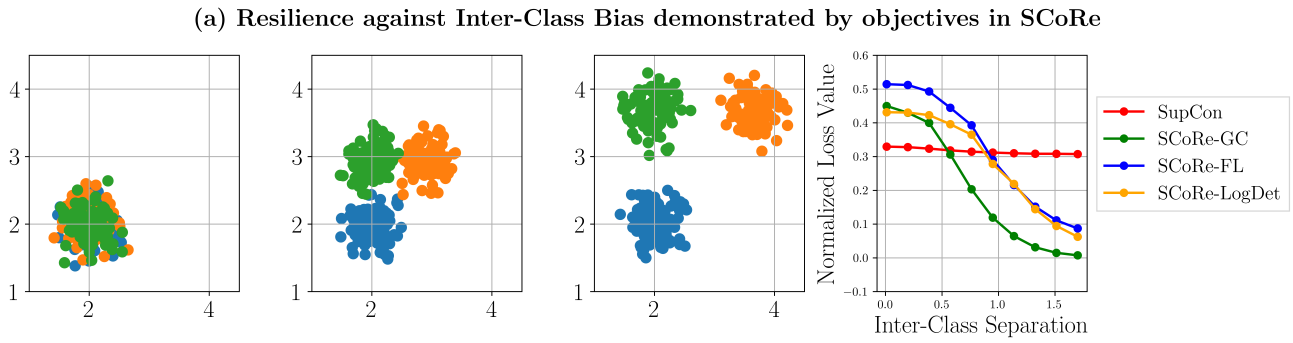
LVIS (Gupta et al., 2019) dataset depicts an extreme case of longtail imbalance with a large number of tail classes. The dataset consists of 1203 classes created by extending the label set in MS-COCO (Lin et al., 2014) (consisting of just 80 classes). We adopt the version v1.0 of LVIS for our experiments and conduct our experiments on Faster-RCNN+FPN architecture adopting a ResNet-101 backbone with a batch size of 16, initial learning rate of 0.06 and repeat factor sampling for a total of 180,000 iterations. Similar to IDD , the step based learning rate scheduler is adopted to reduce the learning rate by 1/10 at regular intervals.

A.3. Experiments on Synthetic Datasets

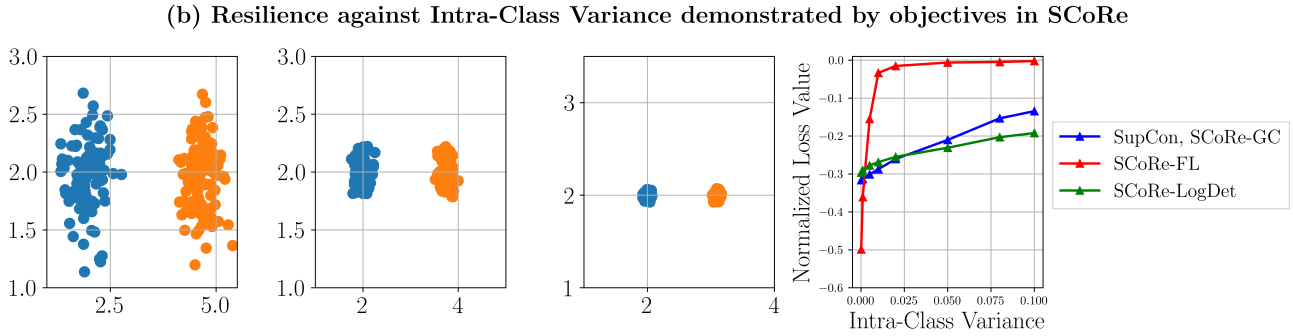
The aim of the experiments on synthetic datasets in SCoRe is to demonstrate the variation of SCoRe objectives to - *inter-class bias and intra-class variance* and *class-imbalance in long-tail settings*.

To demonstrate resilience to inter-class bias we vary the inter-cluster separation between three disjoint clusters with constant variance of 0.05 as shown in Figure 9. In Figure 9(a) we plot the variation of the losses introduced in Table 1 to inter-class bias arising to increased overlaps between feature clusters. **We observe a large variation to inter-class bias in SCoRe objectives over existing SoTA methods** demonstrating their resilience to the phenomenon (since the loss would be minimized during back-propagation).

To demonstrate resilience to intra-class variance, we keep the inter-cluster distance (separation between the extremities



(a) Inter-Class Bias



(b) Intra-Class Variance

Figure 9. **Resilience of instantiations in SCoRe towards Intra-Class Variance (a) and Inter-Class Bias (b) in representation learning tasks demonstrated through pathologically created synthetic benchmarks.**

of two clusters) between two clusters to be constant and vary the variance within each cluster in the range of [0.01, 0.1]. Similar to above we plot the variation of losses in Table 1 through Figure 9(b). The plots show that SCoRe objectives, especially SCoRe-FL **demonstrates a relatively large variation to increase in intra-class variance over existing SoTA approaches** like SupCon (Khosla et al., 2020) establishing combinatorial objectives to be resilient to the phenomenon.

Note, that for our experiments all synthetic data-points lie in the first quadrant and thus we adopt the Radial Bias Field (RBF) kernel (Killamsetty et al., 2021) as the similarity metric for computing S_{ij}.

A.4. Ablation Study : SCoRe Objectives Learn Discriminative Features with Fewer Training Rounds

In this experiment we compare the number of training epochs in stage 1 required to learn discriminative feature representations for various instances of SCoRe against SoTA contrastive learner (SupCon). In stage 1 we train our models (with varying objectives) for various number of epochs in the range of [100, 1000] on the balanced and imbalanced settings of the CIFAR-10 dataset. We keep the remaining hyperparameters constant by training our models with a initial learning rate of 0.4, batch size of 512, temperature 0.7 and a cosine annealing scheduler. For each of these trained models from stage 1 we train the classifier Clf in stage 2 until convergence and report the Top-1 Accuracy %. Although, the performance (Top-1 Accuracy %) of our models in balanced setting is lower than the SoTA contrastive learners we observe in Figure 10 that combinatorial objectives like SCoRe-FL learn robust representations within very few training rounds in stage 1 of model training as compared to SoTA approach SupCon (Khosla et al., 2020). Interestingly, in

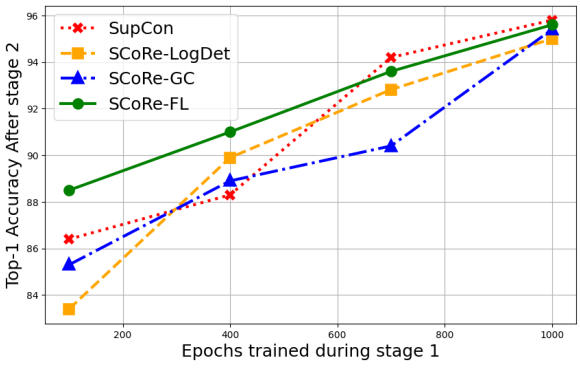


Figure 10. **Ablation Study**: SCoRe demonstrates faster learning of discriminative representations even on the *balanced* setting of CIFAR-10 dataset.

the imbalanced setting as shown in Figure 6 models trained with combinatorial objectives in SCoRe (SCoRe-FL, SCoRe-GC and SCoRe-LogDet) learn better discriminative features within fewer number of epochs in stage 1. This allows model developers to rapidly train generalizable feature extractors for several downstream applications.

A.5. Ablation Study : Effect of λ on performance of Graph-Cut based Objective

In this section we perform experiments on the hyperparameter λ introduced in Graph-Cut based combinatorial objective in SCoRe. The hyper-parameter λ is applied to the sum over the penalty associated with the positive set forming tighter clusters. This parameter controls the degree of compactness of the feature cluster ensuring sufficient diversity is maintained in the feature space. For SCoRe-GC to be submodular it is also important for $\lambda$to be greater than or equal to 1 ($\lambda \geq 1$). For this experiment we train the two stage framework in SCoRe on the longtail CIFAR-10 dataset for 500 epochs in stage 1 with varying λ values in range of [0.5, 2.0] and report the top-1 accuracy after stage 2 model training on the validation set of CIFAR-10. Table 7 shows that we achieve highest performance for $\lambda = 1$ for longtail image classification task on the CIFAR-10 dataset. We adopt this value for all experiments conducted on GC in this paper.

Table 7. **Ablation study** on the effect of λ on the performance of Graph-Cut based combinatorial objective in SCoRe.

λ	Top-1 acc CIFAR-10 (longtail)
0.5	83.65
1.0	**89.96**
1.5	87.11
2.0	85.86

A.6. Proof of theorems of Submodular Combinatorial Objectives

A.6.1. Facility Location

In this section we show proofs for the introduced L_{S_f} and L_{C_f} for the facility location function in Theorem 3.1.

Proof. The facility location function over a set A can be represented as $f(A) = \sum_{i \in \mathcal{V}} \max_{j \in A} S_{ij}$. Theorem 3.1 instantiates this function in SCoRe to define two combinatorial objectives L_{S_f} and L_{C_f} as presented by Equation (2) in the main paper.

From the definition of L_{S_f}, the SCoRe-FL (L_{S_f}) objective can be derived as $L_{S_f}(\theta) = \sum_{k=1}^{|C|} \frac{1}{|\mathcal{V}|} f(A_k, \theta)$. Substituting the instance of FL $f(A_k, \theta) = \sum_{i \in \mathcal{V}} \max_{j \in A_k} S_{ij}$ in the equation we get:

$$
\begin{aligned}
L_{S_f}(\theta) &= \frac{1}{|\mathcal{V}|} \sum_{k=1}^{|C|} f(A_k, \theta) \\
&= \frac{1}{|\mathcal{V}|} \sum_{k=1}^{|C|} \sum_{i \in \mathcal{V}} \max_{j \in A_k} S_{ij} \\
&= \frac{1}{|\mathcal{V}|} \sum_{k=1}^{|C|} \sum_{i \in \mathcal{V} \setminus A_k} \max_{j \in A_k} S_{ij} + \sum_{k=1}^{|C|} \sum_{i \in A_k} \max_{j \in A_k} S_{ij} \\
L_{S_f}(\theta) &= \frac{1}{|\mathcal{V}|} \sum_{k=1}^{|C|} \sum_{i \in \mathcal{V} \setminus A_k} \max_{j \in A_k} S_{ij} + |\mathcal{V}|, \text{ since } \sum_{i \in A_k} \max_{j \in A_k} S_{ij} \text{ is a constant over the set } A_k
\end{aligned}
$$

Now, considering the C_f formulation in Equation (1) and substituting $f(A_k, \theta)$ we get:

$$
\begin{aligned}
L_{C_f}(\theta) &= \frac{1}{|\mathcal{V}|} \left[\sum_{k=1}^{|C|} f(A_k, \theta) - f(\bigcup_{k=1}^{|C|} A_k, \theta) \right] \\
L_{C_f}(\theta) &= L_{S_f}(\theta) - \frac{1}{|\mathcal{V}|} f(\bigcup_{k=1}^{|C|} A_k, \theta)
\end{aligned}
$$

We know that $\bigcup_{k=1}^{|C|} A_k = \mathcal{V}$. Thus,

$$\begin{aligned}
L_{C_f}(\theta) =& \frac{1}{|\mathcal{V}|}\sum_{k=1}^{|C|}\sum_{i\in\mathcal{V}\setminus A_k}\max_{j\in A_k} S_{ij} + |\mathcal{V}| - f(\mathcal{V},\theta) \\
=& \frac{1}{|\mathcal{V}|}\sum_{k=1}^{|C|}\sum_{i\in\mathcal{V}\setminus A_k}\max_{j\in A_k} S_{ij} + |\mathcal{V}| - \sum_{i\in\mathcal{V}}\max_{j\in\mathcal{V}} S_{ij} \\
=& \frac{1}{|\mathcal{V}|}\sum_{k=1}^{|C|}\sum_{i\in\mathcal{V}\setminus A_k}\max_{j\in A_k} S_{ij}, \text{ since } \sum_{i\in\mathcal{V}}\max_{j\in\mathcal{V}} S_{ij} \text{ is a constant over the ground set } \mathcal{V}
\end{aligned}$$

Thus, we show that both L_{S_f} as well as L_{C_f} versions of the Facility-Location (SCoRe-FL) function introduced in Theorem 3.1 are instances of the SCoRe framework. □

A.6.2. GRAPH-CUT

In this section we show proofs for the introduced L_{S_f} and L_{C_f} for the Graph-Cut function in Theorem 3.2.

Proof. The Graph-Cut function over a set A can be represented as $f(A) = \sum_{i\in A, j\in V\setminus A} S_{ij}(\theta) - \lambda\sum_{i,j\in A} S_{ij}(\theta)$. Theorem 3.2 instantiates this function in SCoRe to define two combinatorial objectives L_{S_f} and L_{C_f} as presented by Equation (3) in the main paper.

From the definition of L_{S_f}, the SCoRe-GC (L_{S_f}) objective can be derived as $L_{S_f}(\theta) = \sum_{k=1}^{|C|}\frac{1}{|A_k|} f(A_k,\theta)$. Substituting the instance of GC $f(A_k,\theta)$ in the equation we get:

$$\begin{aligned}
L_{S_f}(\theta) &= \sum_{k=1}^{|C|}\frac{1}{|A_k|} f(A_k,\theta) \\
&= \sum_{k=1}^{|C|}\frac{1}{|A_k|}\sum_{i\in A_k, j\in V} S_{ij}(\theta) - \frac{\lambda}{|A_k|}\sum_{i,j\in A_k} S_{ij}(\theta) \\
&= \sum_{k=1}^{|C|}\frac{1}{|A_k|}\sum_{i\in A_k, j\in V\setminus A_k} S_{ij}(\theta) + \frac{1}{|A_k|}\sum_{k=1}^{|C|}\sum_{i\in A_k, j\in A_k} S_{ij}(\theta) - \frac{\lambda}{|A_k|}\sum_{i,j\in A_k} S_{ij}(\theta)
\end{aligned}$$

Here, the term $\sum_{k=1}^{|C|}\sum_{i\in A_k, j\in A_k} S_{ij}(\theta)$ represents a sum of pairwise similarities over all sets in $\mathcal{V}$. Thus, its value is a constant for a fixed training/ evaluation dataset. Using this condition and ignoring the constant term, we can show that :

$$L_{S_f}(\theta) = \sum_{k=1}^{|C|}\frac{1}{|A_k|}\left[\sum_{i\in A_k, j\in V\setminus A_k} S_{ij}(\theta) - \lambda\sum_{i,j\in A_k} S_{ij}(\theta)\right]$$

Now, considering the C_f formulation in Equation (1) and substituting $f(A_k, \theta)$ we get:

$$L_{C_f}(\theta) = \sum_{k=1}^{|C|} \frac{1}{|A_k|} f(A_k, \theta) - f(\bigcup_{k=1}^{|C|} A_k, \theta)$$

$$L_{C_f}(\theta) = L_{S_f}(\theta) - \frac{1}{|A_k|} f(\bigcup_{k=1}^{|C|} A_k, \theta)$$

We know that $\bigcup_{k=1}^{|C|} A_k = \mathcal{V}$. Thus,

$$L_{C_f}(\theta) = \sum_{k=1}^{|C|} \frac{1}{|A_k|} \left(\sum_{i \in A_k, j \in V \setminus A_k} S_{ij}(\theta) - \lambda \sum_{i,j \in A_k} S_{ij}(\theta) \right) - \frac{1}{|A_k|} \left(\sum_{i \in \bigcup_{k=1}^{|C|} A_k, j \in V \setminus \bigcup_{k=1}^{|C|} A_k} S_{ij}(\theta) - \lambda \sum_{i,j \in \bigcup_{k=1}^{|C|} A_k} S_{ij}(\theta) \right)$$

Since the sets A_k are disjoint, we can simplify the expression as :

$$\sum_{k=1}^{|C|} \frac{1}{|A_k|} \left[\sum_{i,j \in A_k} S_{ij}(\theta) = \sum_{i,j \in \bigcup_{k=1}^{|C|} A_k} S_{ij}(\theta) \right]$$

This leads to a cancellation of the terms involving λ in $L_{C_f}(\theta)$:

$$L_{C_f}(\theta) = \sum_{k=1}^{|C|} \frac{1}{|A_k|} \sum_{i \in A_k, j \in V \setminus A_k} S_{ij}(\theta)$$

□

Thus, we show that both L_{S_f} as well as L_{C_f} versions of the Graph-Cut (SCoRe-GC) function introduced in Theorem 3.2 are instances of the SCoRe framework. In the final equation shown in Section 3.1.1 we multiply L_{C_f} with a hyper-parameter λ to control the penalization of the cross similarity between sets A_k and $\mathcal{V} \setminus A_k$.

A.6.3. LOG-DETERMINANT

In this section we show proofs for the introduced L_{S_f} and L_{C_f} for the LogDet function in Theorem 3.3.

Proof. The submodular function $f(A_k, \theta)$ for Log-Determinant is denoted as $\log\det(S_{A_k} + \lambda \mathbb{I}_{|\mathcal{A}_\||})$ with $\lambda\mathbb{I}$ as the identity term used for numerical stability (empirically). Substituting $f(A_k, \theta)$ in the L_{S_f} formulation in Equation (1) we get:

$$L_{S_f}(\theta) = \sum_{k=1}^{|C|} \frac{1}{|A_k|} f(A_k, \theta)$$

$$= \sum_{k=1}^{|C|} \frac{1}{|A_k|} \log\det(S_{A_k}(\theta) + \lambda \mathbb{I}_{|A_k|})$$

This leads to the S_f formulation of the LogDet objective in SCoRe. Now, considering the C_f formulation in Equation (1) and substituting $f(A_k, \theta)$ we get:

$$L_{C_f}(\theta) = \sum_{k=1}^{|C|} \frac{1}{|A_k|} f(A_k, \theta) - f(\bigcup_{k=1}^{|C|} A_k, \theta)$$

$$L_{C_f}(\theta) = L_{S_f}(\theta) - \frac{1}{|A_k|} f(\bigcup_{k=1}^{|C|} A_k, \theta)$$

We know that $\bigcup_{k=1}^{|C|} A_k = \mathcal{V}$. Thus,

$$\begin{aligned} L_{C_f}(\theta) =& \sum_{k=1}^{|C|} \frac{1}{|A_k|} \log\det(S_{A_k}(\theta) + \lambda \mathbb{I}_{|A_k|}) - \frac{1}{|A_k|} f(\mathcal{V}, \theta) \\ =& L_{S_f}(\theta) - \frac{1}{|A_k|} \log\det(S_{\mathcal{V}}(\theta) + \lambda \mathbb{I}_{|\mathcal{V}|}) \\ =& \sum_{k=1}^{|C|} \frac{1}{|A_k|} \left[\log\det(S_{A_k}(\theta) + \lambda \mathbb{I}_{|A_k|}) - \log\det(S_{\mathcal{V}}(\theta) + \lambda \mathbb{I}_{|\mathcal{V}|}) \right] \end{aligned}$$

Thus, we show that both LS_f as well as L_{C_f} versions of the Log-Determinant (LogDet) function introduced in Theorem 3.3 are instances of the SCoRe framework. □

A.7. Proof of Submodularity for Existing Metric/Contrastive Learners

In this section we discuss in depth the submodular counterparts of three existing objective functions in contrastive learning. We provide proofs that these functions are non-submodular in their existing forms and can be reformulated as submodular objectives through modifications without changing the characteristics of the loss function.

A.7.1. Triplet loss and Submod-Triplet loss

Theorem A.1. *The Triplet loss $L(\theta)$, depicted in row 2 of Table 1 is not an instance of SCoRe in its original form while the slightly modified form, Submod-Triplet $L(\theta) = \sum_{k=1}^{|C|} \frac{1}{|A_k|} \sum_{\substack{i \in A_k \\ n \in \mathcal{V} \setminus A_k}} S_{in}^2(\theta) - \sum_{i,p \in A} S_{ip}^2(\theta)$ is an instance of SCoRe (S_f version) with the submodular function $f(A, \theta) = \sum_{\substack{i \in A \\ n \in \mathcal{V} \setminus A}} S_{in}^2(\theta) - \sum_{i,p \in A} S_{ip}^2(\theta)$, defined over a set A.*

Triplet Loss : Here we provide the proof of non-submodularity for the Triplet loss discussed in Theorem A.1.

Proof. We first show that the Triplet loss is not necessarily submodular. The reason for this is the Triplet loss is of the form: $\sum_{n \in \mathcal{V}} \sum_{i,p \in A} D_{in} - \sum_{i,p \in A} D_{ip}$. Note that this is actually supermodular since $-\sum_{i,p \in A} D_{ip}$ is submodular and $\sum_{i,p,n \in A} D_{ip}$ is submodular. As a result, the Triplet loss is **not necessarily submodular**. □

Submod-Triplet : Here we provide the proof for the Submod-Triplet loss in Theorem A.1.

Proof. Submodular Triplet loss (Submod-Triplet) is exactly the same as Graph-Cut where we use $\lambda = 1$ and the similarity as the squared similarity function. Thus, this function is **submodular** in nature. □

A.7.2. Soft-Nearest Neighbor (SNN) loss and Submod-SNN loss

Theorem A.2. *The Soft-Nearest Neighbor (SNN) loss $L(\theta)$, depicted in row 3 of Table 1 is not submodular in its original form while the slightly modified form, Submod-SNN loss $L(\theta) = \sum_{k=1}^{|C|} \frac{1}{|A_k|} \sum_{i \in A_k} [\log \sum_{j \in A_k} \exp(D_{ij}(\theta)) + \log \sum_{j \in \mathcal{V} \setminus A_k} \exp(S_{ij}(\theta))]$ is an instance of SCoRe (S_f version) with the submodular function $f(A_k; \theta) = \sum_{i \in A_k} [\log \sum_{j \in A_k} \exp(D_{ij}(\theta)) + \log \sum_{j \in \mathcal{V} \setminus A_k} \exp(S_{ij}(\theta))]$.*

SNN Loss: Here we provide the proof of non-submodularity for the SNN loss discussed in Theorem A.2.

Proof. From the set representation of the SNN loss we can describe the objective $L(\theta)$ as in Equation 5 . This objective function can be split into two distinct terms labelled as *Term 1* and *Term 2* in the equation above.

$$L(\theta) = \sum_{k=1}^{|C|} - \underbrace{\sum_{i \in A_k} [\log \sum_{j \in A_k} \exp(S_{ij}(\theta))}_{\text{Term 1}} \underbrace{- \log \sum_{j \in \mathcal{V} \setminus A_k} \exp(S_{ij}(\theta))]}_{\text{Term 2}} \tag{5}$$

We prove the objective to be submodular by considering two popular assumptions :
(1) The sum of submodular function over a set of classes A_i, $i \in C$, the resultant is submodular in nature.
(2) The concave over a modular function is submodular in nature.
To prove that $L(\theta, A_k)$ is submodular in nature it is enough to show the individual terms (Term 1 and 2) to be submodular. Note that the sum of submodular functions is submodular in nature. Considering $F(A) = \sum_{j \in A_k} S_j$ for any given $i \in A_k$, we see that $\log \sum_{j \in A_k} \exp(D_j(\theta))$ to be modular as it is a sum over terms $\exp(D_j(\theta))$.
We also know from assumption (2) (Fujishige, 2005), that the concave over a modular function is *submodular* in nature, log being a concave function. Thus, $log \sum_{j \in A_k} exp(S_j)$ is submodular function for a given $i \in A_k$. Unfortunately, the negative sum over a submodular function cannot be guaranteed to be submodular in nature. This renders SNN to be **non-submodular** in nature. □

Submod-SNN Loss : Here we provide the proof of submodularity for the submodular function $f(A_k; \theta)$ of the Submod-SNN loss discussed in Theorem A.2. The variation of SNN loss described in Table 1 can be represented as an instance of $f(A_k; \theta)$ as shown in Equation 6. Similar to the set notation of SNN loss we can split the equation into two terms, referred to as *Term 1* and *Term 2* in the equation above.

$$f(A_k; \theta) = \sum_{i \in A_k} [\underbrace{\log \sum_{j \in A_k} \exp(D_{ij}(\theta))}_{\text{Term 1}} + \underbrace{\log \sum_{j \in \mathcal{V} \setminus A_k} \exp(S_{ij}(\theta))}_{\text{Term 2}}] \tag{6}$$

Proof. Considering $F(A) = \sum_{j \in A_k} S_j$ for any given $i \in A_k$, we prove $\log \sum_{j \in A_k} \exp(D_j(\theta))$ to be modular, similar to the case of SNN loss. Further, using assumption (2) mentioned above we prove that the log (a concave function) over a modular function is submodular in nature. Finally, the sum of submodular functions over a set of classes A_k is submodular according to assumption (1). Thus the term 1, $\sum_{i \in A_k} \log \sum_{j \in A_k} \exp(D_{ij}(\theta))$ in the equation of Submod-SNN is proved to be submodular in nature.

The term 2 of the equation represents the total correlation function of Graph-Cut ($L_{C_f}(\theta)$). Since graph-cut function has already been proven to be submodular in (Fujishige, 2005; Iyer et al., 2022) we prove that term 2 is submodular.

Finally, since the sum of submodular functions is submodular in nature, the sum over term 1 and term 2 which constitutes $L(\theta)$ can also be proved to be **submodular**. □

A.7.3. N-Pairs Loss and Orthogonal Projection Loss (OPL)

In Table 1 both N-pairs loss and OPL has been identified to be submodular in nature. In this section we provide proofs of Theorem A.3 and Theorem A.4 to show they are submodular in nature.

N-pairs Loss : The N-pairs loss $L(\theta)$ can be represented in set notation as described in Equation 7 and has been discussed to be submodular according to Theorem A.3.

Theorem A.3. *The N-pairs loss* $L(\theta) = \sum_{k=1}^{|C|} \frac{-1}{|A_k|} [\sum_{i,j \in A_k} S_{ij}(\theta)] + \frac{1}{|A_k|} [\sum_{i \in A_k} log(\sum_{j \in \mathcal{V}} S_{ij}(\theta) - 1)]$ *is an instance of SCoRe (*S_f *version) with the submodular function* $f(A, \theta) = -\sum_{i,j \in A} S_{ij}(\theta) + \sum_{i \in A} log(\sum_{j \in \mathcal{V}} S_{ij}(\theta) - 1)$*, defined over set* A*.*

Proof. Similar to SNN loss, we can split the equation for $f(A_k; \theta)$ into two distinct terms.

$$f(A_k; \theta) = \underbrace{-[\sum_{i,j \in A_k} S_{ij}(\theta)}_{\text{Term 1}} + \underbrace{\sum_{i \in A_k} log(\sum_{j \in \mathcal{V}} S_{ij}(\theta) - 1)]}_{\text{Term 2}} \tag{7}$$

The first term (Term 1) in N-pairs is a negative sum over similarities, which is submodular in nature (Fujishige, 2005). The second term (Term 2) is a log over $\sum_{j \in \mathcal{V}} S_{ij}(\theta) - 1$, which is a constant term for every training iteration as it encompases the whole ground set $\mathcal{V}$. The sum of Term 1 and Term 2 over a set A_k is thus **submodular** in nature. □

OPL : The Orthogonal Projection loss can be represented as Equation 8 in its original form and discussed to be submodular according to Theorem A.4 in the main paper.

Theorem A.4. *The Orthogonal Projection loss (OPL)* $L(\theta) = \sum_{k=1}^{|C|} \frac{1}{|A_k|}(1 - \sum_{i,j\in A_k} S_{ij}(\theta)) + \frac{1}{|A_k|}\sum_{i\in A_k}\sum_{j\in\mathcal{V}\setminus A_k} S_{ij}(\theta)$ *is an instance of SCoRe(S_f version). OPL largely represents the Graph-Cut (GC) function which is also an instance of SCoRe with the submodular function* $f(A,\theta) = (1 - \sum_{i,j\in A} S_{ij}(\theta)) + \sum_{i\in A}\sum_{j\in\mathcal{V}\setminus A} S_{ij}(\theta)$, *defined over a set* A *with* $N_f(A_k) = |A_k|$.

Proof. Similar to above objectives we split the equation for $f(A_k;\theta)$ into two distinct terms and individually prove them to be submodular in nature.

$$f(A_k;\theta) = \underbrace{(1 - \sum_{i,j\in A_k} S_{ij}(\theta))}_{\text{Term 1}} + \underbrace{(\sum_{i\in A_k}\sum_{j\in\mathcal{V}\setminus A_k} S_{ij}(\theta))}_{\text{Term 2}} \tag{8}$$

The Term 1 represents a negative sum over similarities in set A_k and is thus submodular in nature. The Term 2 is exactly L_{C_f} of Graph-Cut (GC) with $\lambda = 1$ and is also submodular in nature. Since the sum of two submodular functions is also submodular, the underlying function $f(A_k;\theta)$ in $L(\theta, A_k)$ as in Equation (8) is also **submodular**. □

A.7.4. Supervised Contrastive (SupCon) Loss

Theorem A.5. *The SupCon loss* $L(\theta) = \sum_{k=1}^{|C|}[\frac{-1}{|A_k|}\sum_{i,j\in A_k} S_{ij}(\theta)] + \sum_{i\in A_k}\frac{1}{|A_k|}\log(\sum_{j\in V} exp(S_{ij}(\theta)) - 1)$ *depicted in row 4 of Table 1 is an instance of SCoRe (S_f version) with the submodular function* $f(A_k,\theta) = -[\sum_{i,j\in A_k} S_{ij}(\theta)] + \sum_{i\in A_k}[log(\sum_{j\in\mathcal{V}\setminus A_k}\exp(S_{ij}(\theta)) - 1)]$, *defined over a set* A *and normalization factor* $N_f(A_k) = |A_k|$.

The combinatorial formulation of SupCon as in Equation 9 can be defined as a sum over the set-function $L(\theta, A_k)$ as described in Theorem A.5 of the main paper.

$$f(A_k;\theta) = \underbrace{-\sum_{i,j\in A_k} S_{ij}(\theta)}_{\text{Term 1}} + \underbrace{\sum_{i\in A_k}\log(\sum_{j\in V} exp(S_{ij}(\theta)) - 1)}_{\text{Term 2}} \tag{9}$$

Proof. The Term 1 of SupCon is a negative sum over similarities of set A_k and is thus submodular. The Term 2 contains a sum of the exponent of similarities ($\sum_{j\in V} exp(S_{ij}(\theta)) - 1$) which is a modular term as the sum is computed over the complete ground set $\mathcal{V}$. The logarithm over this term constituting the complete inter-class term represents a concave over modular function which is submodular in nature. Thus, the underlying function $f(A_k;\theta)$ for the SupCon loss $L(\theta)$ represented in Theorem A.5 is **submodular** in nature.

□

A.7.5. SupCon-Var Loss

Theorem A.6. *The SupCon-Var loss* $L(\theta) = \sum_{k=1}^{|C|}\frac{-1}{|A_k|}[\sum_{i,j\in A_k} S_{ij}(\theta)] + \frac{1}{|A_k|}\sum_{i\in A_k}[log(\sum_{j\in\mathcal{V}\setminus A_k}\exp(S_{ij}(\theta)))]$ *is an instance of SCoRe (S_f version) with the submodular function* $f(A_k,\theta) = -[\sum_{i,j\in A_k} S_{ij}(\theta)] + \sum_{i\in A_k}[log(\sum_{j\in\mathcal{V}\setminus A_k}\exp(S_{ij}(\theta)))]$, *defined over a set* A_k, *for* $k \in [1, C]$ *and normalization factor* $N_f(A_k) = |A_k|$.

The submodular variant of SupCon (SupCon-Var) as shown in Equation 10 can be split into two terms indicated as Term 1 and Term 2.

$$f(A_k;\theta) = \underbrace{-[\sum_{i,j\in A_k} S_{ij}(\theta)]}_{\text{Term 1}} + \underbrace{\sum_{i\in A_k}[log(\sum_{j\in\mathcal{V}\setminus A_k}\exp(S_{ij}(\theta)))]}_{\text{Term 2}} \tag{10}$$

Proof. The Term 1 of the function $f(A_k;\theta)$ in SupCon-Var is a negative sum over similarities of set A_k and is thus submodular. The Term 2 of the equation is also submodular as it is a concave over the modular term $\sum_{j\in\mathcal{V}\setminus A_k}\exp(S_{ij}(\theta)$, with $\log$ being a concave function. Thus, the L_{S_f} form of SupCon-Var is also **submodular** represented as the sum of two submodular functions is submodular in nature. □