IEEE Transactions on Machine Learning in Communications and Networking

Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/TMLCN.2022.1234567

Learning on Bandwidth Constrained Multi-Source Data with MIMO-inspired DPP MAP Inference

Xiwen Chen[1], Huayu Li[2], Rahul Amin[3], Abolfazl Razi[1] Senior Member, IEEE

[1]School of Computing, Clemson University, Clemson, SC, USA
[2] Department of Electrical & Computer Engineering, University of Arizona, Tucson, AZ, USA
[3]Tactical Networks Group, MIT Lincoln Laboratory, Lexington, MA, USA

This material is based upon the work supported by the National Science Foundation under Grant Numbers 2008784 and 2204721.

ABSTRACT

Determinantal Point Process (DPP) is a powerful technique to enhance data diversity by promoting the repulsion of similar elements in the selected samples. Particularly, DPP-based *Maximum A Posteriori* (MAP) inference is used to identify subsets with the highest diversity. However, a commonly adopted presumption of all data samples being available at one point hinders its applicability to real-world scenarios where data samples are distributed across distinct sources with intermittent and bandwidth-limited connections. This paper proposes a distributed version of DPP inference to enhance multi-source data diversification under limited communication budgets. First, we convert the lower bound of the diversity-maximized distributed sample selection from matrix determinant optimization to a simpler form of the sum of individual terms. Next, a determinant-preserved sparse representation of selected samples is formed by the sink as a surrogate for collected samples and sent back to sources as lightweight messages to eliminate the need for raw data exchange. Our approach is inspired by the channel orthogonalization process of Multiple-Input Multiple-Output (MIMO) systems based on the Channel State Information (CSI). Extensive experiments verify the superiority of our scalable method over the most commonly used data selection methods, including GreeDi, Greedymax, random selection, and stratified sampling by a substantial gain of at least 12% reduction in Relative Diversity Error (RDE). This enhanced diversity translates to a substantial improvement in the performance of various downstream learning tasks, including multi-level classification (2%-4% gain in accuracy), object detection (2% gain in mAP), and multiple-instance learning (1.3% gain in AUC).

INDEX TERMS Determinantal Point Process, Data Diversification, Distributed Learning, Distributed Sources.

I. Introduction

Many AI platforms in modern smart city applications, such as Smart Transportation Systems (STS) [1], [2], AI-based Energy Management Systems (EMS) [3], Smart Healthcare Systems (SHS) [4], [5], AI-enabled Live Event Monitoring Systems (LEMS) [6], and Smart Budget Allocation (SBA) [7] rely on data-driven methodologies for service provisioning. The essence of such platforms is exploiting learnable contextual patterns from accumulated data from distinct and often geographically distributed data sources. Nevertheless, these systems are typically constrained in terms of communication bandwidth and storage capacity. A significant source of inefficiency is collecting raw data indiscriminately from these sources. To reduce transmission resource overuse, one may benefit from more refined and selective data pooling, in addition to deploying efficient infrastructure, such as intelligent reflecting surface [8], the ambient backscatter [9], and wireless power transfer technologies [10].

Several solutions have been proposed to improve data accumulation efficiency under constrained communication and computation power, from different and somewhat complementary perspectives. These methods include Data Compression [11], Semantic Communication (SC) [12], and Edge Computing (EC) [13]–[15], which aim to minimize or fully eliminate raw data exchange in one form or another while not compromising the ultimate quality of service for learning-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

based applications. Although these methods substantially enhance the efficiency of network-based learning, determining the most effective data collection strategy remains a legitimate challenge.

Our key contribution is developing a formal way of distributed diversity-maximizing data selection policy to improve the learning quality of downstream learning tasks without allowing full convention among sources. The collected samples are utilized as a training dataset by the center to train a library of Machine Learning (ML) models for service provisioning via performing inference on incoming and unseen data from users.

While adopting random selection strategies can be relieving, it is often suboptimal in many cases. This is because random selection does not consider the inherent relations between data points, particularly the potential overlaps in the feature space, which may result in the failure of a subset in accurately representing the entire dataset, especially when the collected samples are limited. Indeed, it has been known for decades that the *diversity* of selected samples can dramatically enhance the quality of learning applications [16]–[19]. Therefore, selecting data samples that closely mimic the geometrical distribution of the entire dataset (for FL-based applications, this can be the distribution of gradient information) by maximizing cross-sample distances in the original or transformed domain can be advantageous. From the statistical learning theory perspective, diverse data enables the reduction of generalization errors through minimizing the empirical risk during the training phase (i.e., training error). In other words, diverse training data ensures that this approximation is as close as possible to the true risk in the underlying distribution. Additionally, diverse training data can preserve a similar hypothesis space (i.e., a set of possible models) with the entire dataset, leading to a higher probability of obtaining a more generalizable model compared to models trained on less diverse data with limited hypothesis space [20].

Due to its simple form, high interpretability, and high efficiency, *Determinantal Point Processes* (DPP) is commonly used for diversity maximization. DPP can also be used as a probabilistic approach to generate diverse data points. In contrast to other point processing methods, such as Poisson point processing, DPP can be formed solely based on the correlation among elements. It assigns a high probability to the measurement of sets of data points with low similarity, making it a valuable tool for tasks like dimensionality reduction and representative sample selection from large datasets [21], [22]. Additionally, by leveraging the properties of linear algebra, DPP can be used to effectively select subsets from a given dataset [23]–[25].

When data exchange, storage, and processing capacity of learning systems are constrained, we often desire to identify and select the most diverse subset. This goal can be implemented through DPP-based *Maximum A Posteriori* (MAP). Recent studies have implemented a centralized version of this algorithm, where all samples are available in the same location or sources are allowed to share their information with no constraint [22], [26], [27]. However, in most practical systems, data samples are generated by sources located at different positions, where cross-source communication is often infeasible, prohibited, or costly. To mimic such limitations, we also presume band-limited communication between data sources and the processing center, which translates to strict limits on the number of accumulated samples, as considered in the system model in Section IV. Specifically, we assume neither the sources nor the coordinator has global knowledge about the collected samples. Therefore, a conventional DPP MAP inference [22] by traversing all data samples is infeasible, and using some sort of distributed implementation is unavoidable. Perhaps, the most popular approach to distributed DDP inference is a multi-stage method proposed in [28], which first implements a local greedy search by each source to collect candidate samples that are diverse within that source regardless of other sources' samples, and then performs another selection on the accumulated candidates to obtain the final set of samples. This method is suboptimal because the original selection neglects global diversity, as will be presented in our comparative results. Furthermore, this protocol involves sending candidate samples to a central unit, some of which are ultimately discarded in the second stage. In contrast, we impose a zero-communication overhead policy by merely sending the selected samples.

In this paper, drawing inspiration from specific techniques in Multiple-Input Multiple-Output (MIMO) transmission [29]–[31] — particularly those relating to power optimization and pre-coding processes based on Channel State Information (CSI) — we propose an effective and scalable scheduling strategy. It is noteworthy that this strategy is not designed to enhance existing MIMO techniques, rather it borrows ideas from MIMO systems to implement similar techniques to implement a distributed version of diversity-maximizing data collection applicable to a wide range of applications with arbitrary communication systems. The only requirement is the presence of a feedback channel from the sink to data sources that encompasses almost all modern communication systems. The key idea is developing a lightweight feedback mechanism to eliminate the need for sharing actual samples among sources to facilitate global diversity assessment (as shown in Fig. 1). For instance, the total MIMO capacity expression is decomposed to the sum of disjoint individual capacity terms to simplify power optimization (decomposing $\log\det()$ in Eq. 3). We use a similar methodology to break down the global diversity measure into quantifiable terms, each of which depends only on the samples of one source (Theorem 1). An inherent assumption of our approach is interval-by-interval transmission. This approach enables us to design a feedback mechanism to send surrogate diversity measures to each data source. With this feedback, each source can adjust its local selection strategy to select diversity-maximizing samples in a global sense without having access

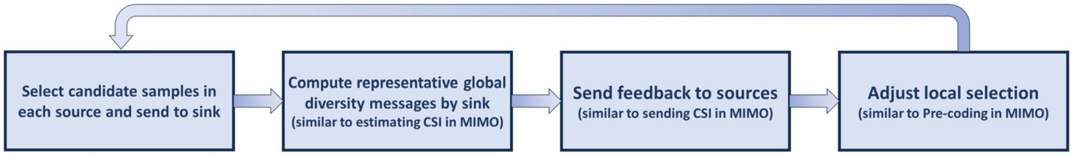

Select candidate samples in each source and send to sink
Compute representative global diversity messages by sink (similar to estimating CSI in MIMO)
Send feedback to sources (similar to sending CSI in MIMO)
Adjust local selection (similar to Pre-coding in MIMO)

Figure 1. The workflow of the proposed MIMO-like distributed diverse data selection framework.

to the samples collected by other sources. These two steps are analogous to *Channel State Information* (CSI) estimation and *Pre-coding*, respectively.

Contributions. We propose a MIMO-inspired strategy for performing the MAP inference on distributed data under communication constraints with the following steps. First, we reformulate the lower bound of the global diversity, which allows us to decompose it into a sum of factors, where each of them can be quantified locally from its source. Afterward, we show that the feedback mechanism can improve the lower bound of diversity (we name it *conditional lower bound*). This tighter lower bound further enhances the diversity of selected samples. Additionally, to address the bandwidth-limited transmission, we propose a determinant-preserving approximation of the feedback messages based on Cauchy–Binet's formula to achieve near-optimal global diversity without sharing data samples. Finally, we investigate the practical benefits of our distributed data selection by evaluating the performance of downstream learning tasks in multiple applications, including Classification, Object Detection, and Multiple-Instance Learning (MIL).

II. Related Work

In this section, we delve into the existing methods to improve data accumulation efficiency from different perspectives, including Data Compression, Semantic Communication, and Distributed and Federated Learning.

Data Compression along with Individual and Distributed Source Coding [32], [33], and Compressive Sampling methods [34] aim to reduce the accumulated data size by exploiting temporal and spatial correlations and sparsity patterns to design more efficient source encoders while retaining essential information to minimize the storage and communication overhead. Conventional compression methods led to the design of efficient encoders for common data modalities (such as MP3/AAC format for audio, JPEG/HEIF for images, and H.264/H.265 for video). Compressive sensing on the other hand intends to take samples far below the Nyquist rate when the signal representation is sparse in a potentially unknown domain (e.g., fMRI imaging [35]). Distributed compression also aims to exploit spatial correlations for joint recovery of data collected from distinct sources (e.g., multi-view imaging [36], [37]). Beyond conventional methods, a recently emerged trend is to harness the astonishing power of Deep Learning (DL) architectures to implement learning-based data encoder-decoder methods for image and video compression, demonstrating enhanced compression ratios [38]–[41].

Another fundamental paradigm shift is departing from content delivery, namely transmitting raw and compressed data batches towards developing Semantic Communication, a knowledge-based approach to convey the semantic content of data to users, especially for learning-based applications. As an illustration, in a traffic safety monitoring system, instead of transmitting complete video frames captured by roadside units, a set of representative features gauging the overall safety of traffic on the road can be sent to the control station [1], [42]. Likewise, Semantic Communication can be deployed by wireless vehicular networks [43] to enable efficient service provisioning for multiple users in vehicle-to-vehicle networks without sharing high-throughput raw imagery. This method involves constructing Knowledge Bases (KBs) that facilitate the extraction and interpretation of semantic information by the sender and receiver, respectively [12]. Such methods well integrate with Edge Computing architectures, where the bulk of the processing is pushed to the network edge in the proximity of data origination sources [44].

Another avenue to solving this issue is using Distributed and Federated Learning (DL/FL), an increasingly embraced approach. FL substitutes data exchange with model-sharing strategies, orchestrating locally constructed models to form unified learning models without the need for sharing massive information. While reducing communication costs, it also mitigates data privacy concerns, especially when equipped with privacy-preserving calculations [13], [45]–[47] and encryption methods [48]. FL has found a particularly warm reception in medical applications, where patient privacy is of paramount concern. Similar to semantic communication, FL can also hugely benefit from the expanding capabilities of ever-growing EC platforms [49].

Our method is applicable to central processing methods where data aggregation is an integral part. It is also applicable to FL, which involves some sort of data delivery to local processors. Further, the global learning quality metrics can be enhanced by selecting the most diverse samples across local models. Our method also well integrates with semantic computation, if proper similarity kernels are designed to characterize the semantic diversity of shared content. In short, our discern data selection method, does not replace, but complements the modern EC, FL, and SC approaches.

III. Background Knowledge

A. Determinant Point Processing (DPP)

DPP is a probability measure defined over $2^{|\mathcal{S}|}$ subsets of $\mathcal{S}$, where $|\mathcal{S}|$ denotes the cardinality of the set $\mathcal{S}$. Suppose a finite dataset is represented by $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n]^\top \in \mathbb{R}^{n\times m}$.

Here, $\mathbf{z}_i$ is a $m \times 1$ column vector representing the i^{th} data sample. Given a *Gram Matrix* $\mathbf{L} = \mathbf{Z}\mathbf{Z}^\top$, *L-ensemble DPP* is presented by having an arbitrary subset A drawn from the entire set $\mathcal{S}$ to satisfy, $\mathcal{P}(A) \propto \det(\mathbf{L}_A)$, where $\mathcal{P}(A)$ denotes the probability of selecting subset A from the entire set $\mathcal{S}$ and $\mathbf{L}_A$ denotes the submatrix of $\mathbf{L}$ with rows and columns indexed by set A. The MAP inference for K-DPP is formulated as,

$$\arg\max_{A \subseteq S} \det(\mathbf{L}_A), \quad s.t.\ |A| = k_T, \tag{1}$$

where A denotes the index set of selected samples and constant k_T denotes the given fixed cardinality, and $k_T \leq rank(\mathbf{L})$ is a necessary condition to ensure $\mathbf{L}_A$ is full-rank, and accordingly, its determinant is greater than 0. This is because the rank of a submatrix never exceeds that of the original matrix, i.e., $rank(\mathbf{L}_A) \leq rank(\mathbf{L})$ [50]. If we have $k_T > rank(\mathbf{L})$, the identity $rank(\mathbf{L}_A) < rank(\mathbf{L})$ implies that $rank(\mathbf{L}) < k_T$ meaning that L_A is not full rank. Without this constraint, k_T could be greater than $rank(\mathbf{L}_A)$, leading $\mathbf{L}_A$ not to be full-rank.

From the geometric perspective, $\det(\mathbf{L}_A)$ represents the square of the volume formed by the feature vectors of selected samples, which occurs for orthogonal vectors [21]. Hence, the DPP MAP problem (Eq. 1) is equivalent to orthogonalizing feature vectors, which leads to a better representation of the feature space.

Since MAP inference is an NP-hard problem, one popular solution is using greedy search and formulating the following sub-modular function, $j = \arg\max_{i \in S \setminus A} \log\det\left(\mathbf{L}_{A \cup \{i\}}\right) - \log\det\left(\mathbf{L}_A\right)$, which can give a $(1-1/e)$-approximation of the optimal solution [28]. Here, j denotes the selected index in each round. The current fastest greedy search proposed in [22] is based on the Cholesky decomposition and requires $\mathcal{O}(n^3)$ complexity for initialization and $\mathcal{O}(k_T^2 n)$ to return k_T items. We denote selection by this method with given Gram matrix $\mathbf{L}$ and the set cardinality k_T as $A^* = \text{MAP-DPP}(\mathbf{L}, k_T)$.

B. Multiple-Input Multiple-Output (MIMO) systems

Before delving into the design of our distributed selection strategy for transmission scheduling, let us briefly review the fundamental principles of MIMO systems that inspired us to develop the proposed method. In a MIMO system, a signal vector $\mathbf{s} \in \mathbb{C}^{m_t}$ is transmitted by M antennas $(TX_1, \cdots, TX_{m_t},)$ to be received as a vector $\mathbf{r} \in \mathbb{C}^{m_r}$ by N antennas $(RX_1, \cdots, RX_{m_r})$. The link between a transmitting antenna TX_i and a receiving antenna RX_j is represented by an element $\mathbf{G}_{i,j}$ of the channel matrix $\mathbf{G} \in \mathbb{C}^{m_t \times m_r}$. The channels are influenced by factors such as multi-path fading and interference, causing different link conditions. Mathematically, a MIMO system can be presented as $\mathbf{r} = \mathbf{G}\mathbf{s} + \mathbf{n}$, where $\mathbf{n} \in \mathbb{C}^{m_r}$ is often modeled as an additive zero-mean Gaussian-distributed noise with covariance $\sigma^2\mathbf{I}$, i.e. $\mathbf{n} \sim \mathcal{CN}(0, \sigma^2\mathbf{I})$. Apparently, when the channels are highly correlated and the rank of $\mathbf{G}$ is low, the equations for recovering $\mathbf{s}$ from $\mathbf{r}$ become underdetermined. To mitigate the interference of different antennas at the receiver, *pre-coding* is employed to orthogonalize data between channels by utilizing channel state information (CSI).

We assume $m_t = m_r$ and let $\mathbf{Q} = \mathbb{E}[\mathbf{s}\mathbf{s}^\dagger]$ denote the covariance matrix of $\mathbf{s}$, where $\mathbf{s}^\dagger$ denotes the conjugate transpose of $\mathbf{s}$. Inequality trace$(\mathbf{Q}) \leq \rho$ always holds for preserving the overall power constraint. The capacity of the system measures the maximum amount of information that can be transferred with an arbitrary small error. According to [51], the capacity is:

$$\max_{\mathbf{Q}} C = \log\det(\mathbf{I} + \frac{1}{\sigma^2}\mathbf{G}\mathbf{Q}\mathbf{G}^\dagger) \quad s.t.\ \text{trace}(\mathbf{Q}) \leq \rho. \tag{2}$$

Let SVD of $\mathbf{G}$ be $svd(\mathbf{G}) = \mathbf{U}\mathbf{S}\mathbf{V}^\dagger$, and λ_i represent the i-th singular value of $\mathbf{G}$, which corresponds to the diagonal element of $\mathbf{S}$. Then, the optimal solution of $\mathbf{Q}$ can be expressed as $\mathbf{V}\mathbf{P}\mathbf{V}^\dagger$, where $\mathbf{P}$ is a diagonal matrix with elements $\{p_i\}_{1 \leq i \leq N}$. Then, the problem in Eq. (2) can be reformulated accordingly:

$$\begin{aligned} &\max_{\text{trace}(\mathbf{P}) \leq \rho} \log\det\left(\mathbf{I} + \frac{1}{\sigma^2}\mathbf{U}\mathbf{S}^2\mathbf{P}\mathbf{U}^\dagger\right) \qquad (3) \\ \overset{(a)}{=} &\max_{\text{trace}(\mathbf{P}) \leq \rho} \sum_{i=1}^{N} \log\left(1 + \frac{p_i\lambda_i^2}{\sigma^2}\right). \end{aligned}$$

The equality (a) holds because $\left(1 + \frac{p_i\lambda_i^2}{\sigma^2}\right)$ is the eigenvalue of $\left(\mathbf{I} + \frac{1}{\sigma^2}\mathbf{U}\mathbf{S}^2\mathbf{P}\mathbf{U}^\dagger\right)$, and the product of the all eigenvalues of a square matrix is equal to its determinant [52]. Note that $\log\left(1 + \frac{p_i\lambda_i^2}{\sigma^2}\right)$ is concave in p_i and represents the capacity of the Single-Input Single-Output (SISO) channel with transmission power p_i, channel gain λ_i, and noise variance σ^2. Therefore, using SVD-based pre-coding can translate the original matrix optimization problem in Eq. (2) to a much simpler form of Eq. (3), which can be solved by standard convex optimization algorithms. To realize SVD-based precoding, we use $\widetilde{\mathbf{s}} = \mathbf{V}\mathbf{P}^{1/2}\mathbf{s}$ to orthogonalize the transmitted signal. We will use a similar approach to decompose the global diversity measurement into a sum of individual terms that purely depend on the samples within each source.

IV. System Model

In this work, we will follow a system model presented in Fig. 2, which consists of a processing center, distributed sources, and several distinct data users. This model ensures generalization to different data types, learning models, application domains, and communication systems. Although we considered one target application, one processing center, and distinct data source and users, extension to more complex setups where some nodes have both source and service user functions or systems with multiple processing centers is straightforward. The processing center pools data from distributed sources in an interval-by-interval fashion to build a learning-based data processing model (or a library of models). In each interval, the processing center computes and sends the spare feedback to sources based on the received data. These messages can be used as surrogate measures for the accumulated data to compute and enforce global diversity. Then, each source adjusts its selection strategy based on the received feedback and prepares data for the next round of transmission. The interval-by-interval transmission fully aligns with the current trend of online deep learning methods where training is typically performed for a received data batch, because re-training for a single data sample is known to be inefficient. This continued process enables the processing center to train and maintain a library of Machine Learning (ML) and inference models for seamless service provisioning. Our main objective is to develop optimal data selection strategies by distributed sources with minimal data sharing overhead. It is worth mentioning that the selected samples in the center are used to train the models, and the inference is performed on new, unseen data samples from users. The workflow of our proposed method is presented in Fig. 1. We evaluate the effectiveness of the selection methods from two perspectives: i) the diversity of the selected samples defined by DPP, and ii) more importantly, the learning quality of the downstream tasks, which can be evaluated by task-specific metrics (e.g., accuracy and F1 score for classification and mean average precision (mAP) for object detection). A summary of metrics used in our evaluation is provided in Table 1. Please refer to Section VII and Section VIII for the detailed evaluation.

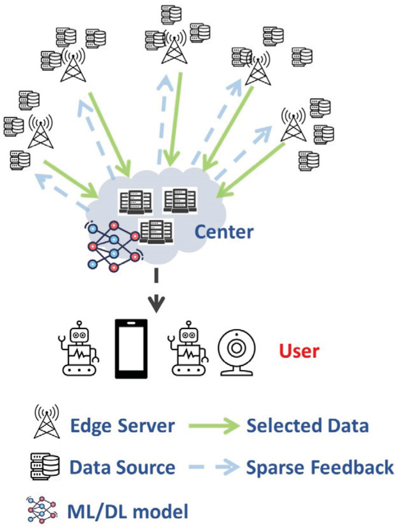


Figure 2. **A typical workflow of data-driven service provisioning systems. In each round, data sources select data samples and transmit them to the center. Data selection is rendered based on feedback messages provided by the center. After data acquisition, the center can train a library of ML/DL models for service provisioning.**

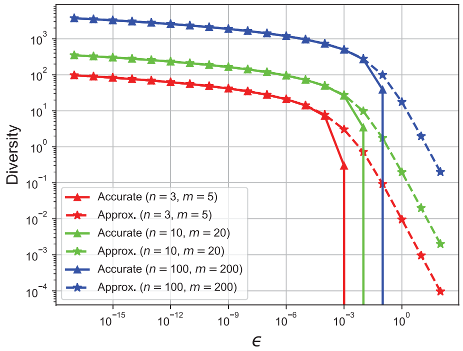


Figure 3. **Illustration of the approximation in computing diversity in logarithmic scale. Here, the *accurate diversity* is computed by $\det(\frac{1}{\epsilon}\mathbf{Z}_{\mathcal{A}}\mathbf{Z}_{\mathcal{A}}^\top)$ (shown with solid line and ▲ marker) and its *approximation* is $\det(\frac{1}{\epsilon}\mathbf{Z}_{\mathcal{A}}\mathbf{Z}_{\mathcal{A}}^\top + \mathbf{I})$ (shown with dashed line and $\star$ marker). We randomly generate a data matrix $\mathbf{Z}_{\mathcal{A}}$, where $(\mathbf{Z}_{\mathcal{A}})_i \sim \mathcal{N}(0, \mathbf{I}_m)$. Parameters n and m denote the number of samples and dimensions of $\mathbf{Z}_{\mathcal{A}}$, respectively. For sufficiently small ϵ, the approximation is accurate.**

V. Problem Formulation

First, we present the notations in Table 2. Suppose there are N data sources with disjointed index sets $S_1, S_2, \cdots, S_i, \cdots, S_N$, and $\mathcal{S} = S_1 \cup S_2 \cup S_i \cup \cdots \cup S_N$ representing the indices of the entire set, and $S_i \cap S_j = \emptyset, \forall i \neq j$. The total number of samples is $n = \sum_{i=1}^{N} n_i$, where $n_i = |S_i|$ denotes the cardinality of the set S_i. Also, let $\mathbf{Z} \in \mathbb{R}^{n \times m}$ be the data matrix of the entire set, where $\mathbf{z}_i$ is a $m \times 1$ data vector. Recall that to maximize diversity, we need to optimize the following problem under communication constraints to select a subset $\mathcal{A}$:

$$\arg\max_{\mathcal{A} \subseteq \mathcal{S}} \ \log\det(\mathbf{L}_{\mathcal{A}}), \quad s.t. \quad |\mathcal{A}| = k_T, \tag{4}$$

where, again, $\mathbf{L}_{\mathcal{A}}$ denotes the columns and rows of $\mathbf{L}$ indexed by $\mathcal{A}$, and k_T is the total number of samples we afford to transmit. Likewise, $X_{A,B}$ denotes a submatrix of X with rows and columns indexed by A and B, respectively. If $A = B$ and X is a square, it can be denoted as X_A, otherwise X_A only denotes the rows of a non-squared matrix X indexed by A. Conventional MAP methods require transmitting all samples to the center to solve this optimization problem, which is impractical or costly when data is distributed across different sources. In these cases, the construction of $\mathbf{L}$ is not feasible. However, $\mathbf{L}_{S_1}, \cdots, \mathbf{L}_{S_N}$ can be easily obtained at different sources, where $\mathbf{L}_{S_i} = \mathbf{Z}_{S_i}\mathbf{Z}_{S_i}^\top$, and the subproblems of the global MAP inference include

$$\arg\max_{A_i \subseteq S_i} \ \log\det\left((\mathbf{L}_{S_i})_{A_i}\right), \quad s.t. \quad |A_i| = k_i, \tag{5}$$

where $(\mathbf{L}_{S_i})_{A_i}$ denotes the matrix $\mathbf{L}_{S_i}$ indexed by A_i. Although these sub-problems can be solved locally to maximize within-source diversity, their resulting samples do not necessarily maximize the global diversity, for not considering the similarity of samples across different sources. On the other hand, sharing full knowledge about samples across sources is costly. To mitigate this issue, we propose a lightweight feedback mechanism in our method as a surrogate measure for the diversity of samples. The problem now is to collectively select A_i for source i, which can be achieved by maximizing of $\det(\mathbf{L}_{\mathcal{A}})$ with $\mathcal{A} = A_1 \cup A_2 \cup A_i \cup \cdots \cup A_N$ subject to the constraint $\sum |A_i| = k_T$ and $A_i \cap A_j = \emptyset, \forall i \neq j$. Following [28], we assume each source transmits the same number of samples, which is denoted as, $k_i = k_T/N$.

Table 1. **The metrics used in our evaluation. ↑ means higher is better, and ↓ means lower is better. The range of all metrics is 0-1.**

Tasks		Metrics	Evaluation
Diversity-based Selection		Relative Diversity Error (RDE) (Eq. 24) (↓)	Section VII
Downstream Learning Tasks	Classification	Accuracy (↑), F1 score (↑)	Section VIII.A
	Object Detection	mean Average Precision (mAP) (↑), F1 score (↑)	Section VIII.B
	Multiple-instance Learning	Accuracy (↑), F1 score (↑), Area Under the Receiver Operating Characteristic Curve (AUC) (↑)	Section VIII.C

Table 2. **Some important notations used in Sections V and VI.**

Notation	Description
A_i	The index set of selected samples from source i.
$\mathcal{A}$	The index set of selected samples from all sources. $\mathcal{A} = A_1 \cup A_2 \cup A_i \cup \cdots \cup A_N$.
$\mathcal{B}$	The index set of selected samples until this moment.
$\mathcal{C}$	The index set of selected columns and rows of $\mathbf{H}_i$.
k_T	The total number of selected samples.
k_i	The number of selected samples from source i.
$[k_i]$	The set $\{1, \cdots, k_i\}$.
$\mathbf{L}$	The Gram matrix of the entire dataset. $\mathbf{L} = \mathbf{Z}\mathbf{Z}^\top$.
$\mathbf{L}_{S_i}$	The Gram matrix of data in source i. $\mathbf{L}_{S_i} = \mathbf{Z}_{S_i}\mathbf{Z}_{S_i}^\top$.
m	The dimensions of the dataset.
$[m]$	The set $\{1, \cdots, m\}$.
$\binom{[m]}{k_i}$	The set of k_i-combinations of $[m]$. i.e. $\{\mathcal{V} \mid \mathcal{V} \subseteq [m], \lvert\mathcal{V}\rvert = k_i\}$. e.g., $\binom{[3]}{2} = \{\{1,2\},\{2,3\},\{1,3\}\}$.
n	The number of samples in the dataset.
n_i	The number of samples in source i.
N	The total number of sources.
r_0	The cardinality of $\mathcal{C}$,. $\lvert\mathcal{C}\rvert = r_0$.
r_1	The number of transmitted singular vectors in Eq. 22.
R	The total tolerable sparsity in transmission.
$\mathcal{S}$	The index set of entire dataset.
S_i	The index set of all samples from source i.
Y_i	Information in the center. The index set of selected samples at this moment that are **not** from source S_i. i.e. $Y_i = \mathcal{B} \setminus A_i$.
$\mathcal{Y}$	Information in the center. $\mathcal{Y} = \{Y_1, \cdots, Y_N\}$.
$\mathbf{H}_i$	Perfect feedback information (analogous to CSI in MIMO). Please refer to Eq. 16.
$\hat{\mathbf{H}}_i$	The sparse approximation of $\mathbf{H}_i$.
$\mathbf{Z}$	The data matrix of the dataset. $\mathbf{Z} \in \mathbb{R}^{n \times m}$.
$\mathbf{Z}_{S_i}$	The data matrix of the source i. $\mathbf{Z}_{S_i} \in \mathbb{R}^{n_i \times m}$.
$\widetilde{\mathbf{Z}}_{S_i}$.	$\mathbf{Z}_{S_i}$ after pre-coding (Eq. 23). $\widetilde{\mathbf{Z}}_{S_i} \in \mathbb{R}^{n_i \times m}$.
$X_{A,B}$	A submatrix of X with rows and columns indexed by A and B, respectively. A and B are some index sets.
X_A	If $A = B$ and X is a square, X_A can be denoted as X_A (i.e. $\mathbf{L}_{A_i}$), otherwise X_A only denotes the rows of a non-squared matrix X indexed by A (i.e. $\mathbf{Z}_{A_i}$).

VI. Methodology

A. MIMO-like Decomposition

Similar to [53], since $\mathbf{L}_\mathcal{A}$ is always a positive-definite Hermitian matrix, we can use the approximation $\det(\mathbf{L}_\mathcal{A}) = \epsilon^{|\mathcal{A}|}\det(\frac{1}{\epsilon}\mathbf{L}_\mathcal{A}) \approx \epsilon^{|\mathcal{A}|}\det(\frac{1}{\epsilon}\mathbf{L}_\mathcal{A}+\mathbf{I})$ for a very small ϵ. Some examples shown in Fig. 3 demonstrate there is negligible approximation error for a very small ϵ. Therefore, we can rewrite the optimization problem as,

$$\arg\max_{\mathcal{A}} \log\det(\frac{1}{\epsilon}\mathbf{L}_\mathcal{A} + \mathbf{I}). \tag{6}$$

This formula can then be presented as

$$\begin{aligned}\mathcal{L}_\epsilon := \log\det(\frac{1}{\epsilon}\mathbf{L}_\mathcal{A} + \mathbf{I}) &= \log\det(\frac{1}{\epsilon}\mathbf{Z}_\mathcal{A}\mathbf{Z}_\mathcal{A}^\top + \mathbf{I}) \\ &\overset{(a)}{=} \log\det(\frac{1}{\epsilon}\mathbf{Z}_\mathcal{A}^\top\mathbf{Z}_\mathcal{A} + \mathbf{I}) \\ &= \log\det(\frac{1}{\epsilon}\sum_i^N \mathbf{Z}_{A_i}^\top\mathbf{Z}_{A_i} + \mathbf{I}).\end{aligned} \tag{7}$$

The validity of (a) in Eq. 7 can be established through SVD decomposition. When $N \ll \frac{1}{\epsilon}$ holds, $1/(N\epsilon)$ is still a large number, and we can use approximation $\det(\mathbf{L}_\mathcal{A}) = (N\epsilon)^{|\mathcal{A}|}\det(\frac{1}{N\epsilon}\mathbf{L}_\mathcal{A}) \approx (N\epsilon)^{|\mathcal{A}|}\det(\frac{1}{N\epsilon}\mathbf{L}_\mathcal{A}+\mathbf{I})$, which enables us to solve the following equation (adopted from Eqs. 6 and 7) instead of directly maximizing $\det(\mathbf{L}_\mathcal{A})$,

$$\arg\max_{A_1,\cdots,A_N} \log\det(\frac{1}{\epsilon}\sum_i^N \mathbf{Z}_{A_i}^\top\mathbf{Z}_{A_i}/N + \mathbf{I}). \tag{8}$$

Here, we have $\mathcal{A} = A_1 \cup A_2 \cup A_i \cup \cdots \cup A_N$ and $A_i \cap A_j = \emptyset, \forall i \neq j$. In summary, from now on, our goal is to maximize the following approximate diversity expression

$$\mathcal{L}_{\epsilon N} := \log\det(\frac{1}{\epsilon}\sum_i^N \mathbf{Z}_{A_i}^\top\mathbf{Z}_{A_i}/N + \mathbf{I}). \tag{9}$$

Theorem 1. *The lower bound of the approximated problem $\mathcal{L}_{\epsilon N}$ is given as,*

$$\mathcal{L}_{\epsilon N} \geq \mathcal{L}^{lower} := \frac{1}{N}\sum_i^N \log\det(\frac{1}{\epsilon}\mathbf{Z}_{A_i}\mathbf{Z}_{A_i}^\top + \mathbf{I}). \tag{10}$$

Proof:
First, we apply the concavity of $f(X) = \log\det X$ for positive definite Hermitian matrices X [50], namely, $\log\det(\alpha X_1 + (1-\alpha)X_2) \geq \alpha\log\det X_1 + (1-\alpha)\log\det X_2$ for $\alpha \in (0,1)$ and positive definite Hermitian matrices X_1, X_2. Therefore, we can easily obtain, $\mathcal{L} \geq \mathcal{L}^{lower} = \frac{1}{N}\sum_i^N \log\det(\frac{1}{\epsilon}\mathbf{Z}_{A_i}^\top\mathbf{Z}_{A_i} + \mathbf{I})$. Then we apply the equality presented in Eq. 7(a) to complete the proof. ■

Remark 2. theorem 1 allows us to decompose the global diversity measurement into the sum of individual local diversity measurements.

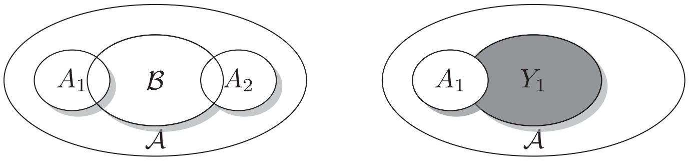


Figure 4. **The Venn Diagram of the set relations. Left: The largest ellipse denotes $\mathcal{A}$. The second largest ellipse denotes $\mathcal{B}$. Two inner ellipses denote the index set A_1 andA_2, respectively. Right: The shadow area denotes Y_1.**

B. MIMO-like CSI Estimation and Pre-coding

Similar to the MIMO systems, where the goal is selecting samples from different sources with minimal interference (through orthogonalization), here we aim to reduce the chance of selecting samples in each source that are similar to the samples of other sources. Ideally, we expect that regardless of the selection by other sources, each source should perform the selection process in parallel so that their collected samples achieve the maximum global diversity, as close as possible. Our approach is block-diagonalization of the similarity matrix $\mathbf{L}$, which is analogous to the orthogonalization process in MIMO. For example, if we achieve full orthogonalization, meaning that all samples of source 1 are orthogonal to all samples of source 2 ($\mathbf{Z}_i\mathbf{Z}_j^\top = 0$ for $\forall i \in S_1, \forall j \in S_2$), then each source can individually maximize the diversity because their samples lie on orthogonal subspaces.

We can achieve this goal by *pre-coding* the samples in each source $\mathbf{Z}_{S_i}$ to $\widetilde{\mathbf{Z}}_{S_i}$ by $\widetilde{\mathbf{Z}}_{S_i} = \mathbf{Z}_{S_i}\mathbf{W}_i$. However, learning $\mathbf{W}_i$ by accessing extra data samples in all sources may not be feasible due to the limited communication budget. Instead, we propose a sparse diversity measurement of selected samples that serves as the pre-coding matrix to tighten the lower bound and guide the selection process.

Recall that $\mathcal{A}$ represents the collection of all selected samples after the completion of the entire selection process. We then denote selected samples until a time point by $\mathcal{B} \subseteq \mathcal{A}$ and define Y_i as the index set of selected samples at this time point that are **not** from source S_i, i.e., $Y_i = \mathcal{B} \setminus (A_i \cap \mathcal{B}) = \mathcal{B} \setminus A_i$. Apparently, we have $A_i \cap Y_i = \emptyset$. It is noteworthy that at that time point, $\mathcal{A}$ is not fully known. Fig. 4 illustrates the relationship of the index sets.

There are two considerations. First, the sources do not drop selected samples because they are used in the next rounds to calculate diversity, but never re-selected for transmission. Second, the feedback message for each source is calculated based on the collected samples from other sources, excluding its own samples to avoid information loss.

To this end, we define $\mathcal{Y} = \{Y_1, \cdots, Y_N\}$, where we have $\mathcal{B} = Y_1 \cup \cdots \cup Y_N$. Therefore, we can rewrite the approximation of the problem in Eq. 7 when conditioned by $\mathcal{Y}$ as

$$\begin{aligned}\mathcal{L}_\epsilon &= \log\det(\frac{1}{\epsilon}\sum_i^N \mathbf{Z}_{A_i}^\top \mathbf{Z}_{A_i} + \mathbf{I}) \qquad (11)\\ &\overset{(a)}{\geq} \mathcal{L}(\cdot \mid \mathcal{Y})\\ &\overset{(b)}{:=} \log\det(\frac{1}{\epsilon}\sum_i^N (\mathbf{Z}_{A_i}^\top \mathbf{Z}_{A_i} + \mathbf{Z}_{Y_i}^\top \mathbf{Z}_{Y_i})/N + \mathbf{I})\\ &\overset{(c)}{=} \log\det(\frac{1}{\epsilon}\sum_i^N (\mathbf{Z}_{\{A_i\cup Y_i\}} \mathbf{Z}_{\{A_i\cup Y_i\}}^\top)/N + \mathbf{I}),\end{aligned}$$

where equality (a) holds when $\mathcal{A} = A_i \cup Y_i = \mathcal{B}$ and (b) defines the conditional problem. The conditional lower bound $\mathcal{L}(\cdot \mid \mathcal{Y})$ can be obtained as,

$$\begin{aligned}\mathcal{L}(\cdot \mid \mathcal{Y}) &\overset{(a)}{\geq} \mathcal{L}^{lower}(\cdot \mid \mathcal{Y}) \qquad (12)\\ &:= \frac{1}{N}\sum_i^N \log\det(\frac{1}{\epsilon}\mathbf{Z}_{A_i}^\top \mathbf{Z}_{A_i} + \frac{1}{\epsilon}\mathbf{Z}_{Y_i}^\top \mathbf{Z}_{Y_i} + \mathbf{I})\\ &\overset{(b)}{=} \frac{1}{N}\sum_i^N \log\det(\frac{1}{\epsilon}\mathbf{Z}_{\{A_i\cup Y_i\}} \mathbf{Z}_{\{A_i\cup Y_i\}}^\top + \mathbf{I}),\end{aligned}$$

where equality (a) holds when $\mathcal{A} = A_i \cup Y_i = \mathcal{B}$. The equation (b) holds is because of $A_i \cap Y_i = \emptyset$.

Theorem 3. *$\mathcal{L}^{lower}(\cdot \mid \mathcal{Y})$ is a tighter bound than $\mathcal{L}^{lower}$ (presented in Theorem 1) to the problem $\mathcal{L}_\epsilon$, which is presented as $\mathcal{L}_\epsilon \geq \mathcal{L}^{lower}(\cdot \mid \mathcal{Y}) \geq \mathcal{L}^{lower}$.*

Proof:
First, $\mathcal{L}_\epsilon \geq \mathcal{L}^{lower}(\cdot \mid \mathcal{Y})$ is proved in Eq. 11. Now, we need to prove $\mathcal{L}^{lower}(\cdot \mid \mathcal{Y}) \geq \mathcal{L}^{lower}$,

$$\begin{aligned}&\frac{1}{N}\sum_i^N \log\det(\frac{1}{\epsilon}\mathbf{Z}_{\{A_i\cup Y_i\}} \mathbf{Z}_{\{A_i\cup Y_i\}}^\top + \mathbf{I}) \qquad (13)\\ &\overset{(a)}{\geq} \frac{1}{N}\sum_i^N \log\det(\frac{1}{\epsilon}\mathbf{Z}_{A_i} \mathbf{Z}_{A_i}^\top + \mathbf{I}).\end{aligned}$$

To this end, we observe LFS and RHS have the same form of a submodular function. Since $A_i \subseteq A_i \cup Y_i$, obviously, the inequality is proved.

An example is shown in Fig. 5, which shows that the conditional lower bound is a tighter bound than the bound presented in Theorem 1.

■

Remark 4. We note that $\mathcal{L}^{lower}(\cdot \mid \mathcal{Y})$ is still a form of the sum of individual terms. Therefore, this theorem demonstrates that, by taking the received information $\mathcal{Y}$ merely from the center as feedback, we can adjust the future selection in each source to enhance the overall diversity.

Remark 5. in fact, $\mathcal{L}^{lower}(\cdot \mid \mathcal{Y})$ can be updated sequentially as$\mathcal{L}^{lower}(\cdot \mid \mathcal{Y}) : \mathcal{L}^{lower}(\cdot \mid \mathcal{Y}^0) \rightarrow \mathcal{L}^{lower}(\cdot \mid \mathcal{Y}^1) \rightarrow \cdots$, if the selection is online. Here, $\mathcal{Y}^t$ is formed after receiving data samples in the first t intervals.

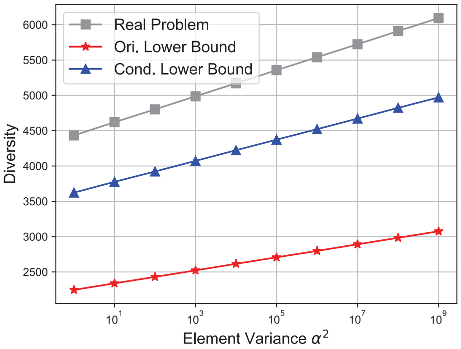


Figure 5. Demonstration of the lower bound and conditional lower bound of the diversity of a two-source scenario, developed in Theorem 3. We generate $\mathbf{Z}_{A_1} \in \mathbb{R}^{80\times 200}$ by $(\mathbf{Z}_{A_1})_i \sim \mathcal{N}(0, \mathbf{I}_{200})$. Similarly, we generate $\mathbf{Z}_{A_2} \in \mathbb{R}^{80\times 200}$ by $(\mathbf{Z}_{A_2})_i \sim \mathcal{N}(0, \alpha^2\mathbf{I}_{200})$. Y_1, Y_2 are randomly selected from $A2, A1$, respectively. The diversity increases with element variance α^2, as expected. The derived condition lower bound is consistently tighter than the original lower bound.

Now, according to Remark 4, the target of each source is switched to perform local selection adjusted by feedback, which is equivalent to maximize $\log\det(\frac{1}{\epsilon}\mathbf{Z}_{\{A_i\cup Y_i\}}\mathbf{Z}_{\{A_i\cup Y_i\}}^\top + \mathbf{I})$ individually in each source. Fortunately, this problem rolls back to a MAP inference as

$$\begin{aligned} &\max_{A_i\subseteq S_i} \det\left(\mathbf{Z}_{\{A_i\cup Y_i\}}\mathbf{Z}_{\{A_i\cup Y_i\}}^\top\right), \\ &s.t. \quad |A_i| = k_i,\ A_i \cap Y_i = \emptyset. \end{aligned} \tag{14}$$

According to Schur's complement, we can re-write it as,

$$\begin{aligned} &\det(\mathbf{Z}_{\{A_i\cup Y_i\}}\mathbf{Z}_{\{A_i\cup Y_i\}}^\top) \\ &= \det(\mathbf{Z}_{Y_i}\mathbf{Z}_{Y_i}^\top)\det\left(\mathbf{Z}_{A_i}\left(\mathbf{I} - \mathbf{Z}_{Y_i}^\top(\mathbf{Z}_{Y_i}\mathbf{Z}_{Y_i}^\top)^{-1}\mathbf{Z}_{Y_i}\right)\mathbf{Z}_{A_i}^\top\right). \end{aligned} \tag{15}$$

Since $\det(\mathbf{Z}_{Y_i}\mathbf{Z}_{Y_i}^\top)$ is fixed, the only information depending on Y_i is

$$\mathbf{H}_i := \left(\mathbf{I} - \mathbf{Z}_{Y_i}^\top(\mathbf{Z}_{Y_i}\mathbf{Z}_{Y_i}^\top)^{-1}\mathbf{Z}_{Y_i}\right). \tag{16}$$

Now, we only need to maximize $\det(\mathbf{Z}_{A_i}\mathbf{H}_i^{1/2}\mathbf{H}_i^{1/2}\mathbf{Z}_{A_i}^\top)$, where $A_i \subseteq S_i$ and $|A_i| = k_i$ still hold. Without loss of information, we can send $\mathbf{H}_i$ (which can be viewed as the CSI) from the center to each source and adjust (which can be viewed as the pre-coding) the data matrix $\mathbf{Z}_{S_i}$ to $\widetilde{\mathbf{Z}}_{S_i} = \mathbf{Z}_{S_i}\mathbf{H}_i^{1/2}$, where $\mathbf{W}_i = \mathbf{H}_i^{1/2}$.

C. Sparse Representation of MIMO-like CSI

To further accommodate the band-limited communication requirements, we seek sending a sparse representation of $\mathbf{H}_i$ than sending the entire information of $\mathbf{H}_i$. To this end, we can obtain the upper bound of the problem,

Theorem 6. *The upper bound of* $\det(\mathbf{Z}_{A_i}\mathbf{H}_i^{1/2}\mathbf{H}_i^{1/2}\mathbf{Z}_{A_i}^\top)$ *is given as*

$$\begin{aligned} &\det(\mathbf{Z}_{A_i}\mathbf{H}_i^{1/2}\mathbf{H}_i^{1/2}\mathbf{Z}_{A_i}^\top) \\ &\leq \det(\mathbf{Z}_{A_i}\mathbf{Z}_{A_i}^\top) \times \sum_{J_1\in\binom{[m]}{k_i}} \det((\mathbf{H}_i)_{J_1}). \end{aligned} \tag{17}$$

Proof:
First, we apply the Cauchy–Binet formula to re-write the optimization problem as,

$$\begin{aligned} &\det(\mathbf{Z}_{A_i}\mathbf{H}_i^{1/2}\mathbf{H}_i^{1/2}\mathbf{Z}_{A_i}^\top) \\ &= \det((\mathbf{Z}_{A_i}\mathbf{H}_i^{1/2})(\mathbf{Z}_{A_i}\mathbf{H}_i^{1/2})^\top) \\ &= \sum_{J_1\in\binom{[m]}{k_i}} \left(\det\left((\mathbf{Z}_{A_i}\mathbf{H}_i^{1/2})_{[k_i],J_1}\right)\right)^2, \end{aligned} \tag{18}$$

where $[m]$ denotes the set $\{1, ..., m\}$ and $\binom{[m]}{k_i}$ denotes the set of k_i-combinations of $[m]$, e.g., $\binom{[3]}{2} = \{\{1,2\},\{2,3\},\{1,3\}\}$. As before, k_i denotes the cardinality of A_i. Then, it is sufficient to prove

$$\left(\det\left((\mathbf{Z}_{A_i}\mathbf{H}_i^{1/2})_{[k_i],J_1}\right)\right)^2 \leq \det(\mathbf{Z}_{A_i}\mathbf{Z}_{A_i}^\top)\det(\mathbf{H}_{J_1}). \tag{19}$$

This is because,

$$\begin{aligned} &\left(\det\left((\mathbf{Z}_{A_i}\mathbf{H}_i^{1/2})_{[k_i],J_1}\right)\right)^2 = \left(\det\left(\mathbf{Z}_{A_i}(\mathbf{H}_i^{1/2})_{[m],J_1}\right)\right)^2 \\ &= \left(\sum_{J_2\in\binom{[m]}{k_i}} \det\left((\mathbf{Z}_{A_i})_{[k_i],J_2}\right)\det\left((\mathbf{H}_i^{1/2})_{J_2,J_1}\right)\right)^2 \\ &\overset{(a)}{\leq} \sum_{J_2\in\binom{[m]}{k_i}} \left(\det\left((\mathbf{Z}_{A_i})_{[k_i],J_2}\right)\right)^2 \\ &\quad \times \sum_{J_2\in\binom{[m]}{k_i}} \left(\det\left((\mathbf{H}_i^{1/2})_{J_2,J_1}\right)\right)^2 \\ &\overset{(b)}{=} \sum_{J_2\in\binom{[m]}{k_i}} \left(\det\left((\mathbf{Z}_{A_i})_{[k_i],J_2}\right)\right)^2 \\ &\quad \times \det\left((\mathbf{H}_i^{1/2})_{J_1,[m]}((\mathbf{H}_i^{1/2})_{J_1,[m]})^\top\right) \\ &\overset{(c)}{=} \sum_{J_2\in\binom{[m]}{k_i}} \left(\det\left((\mathbf{Z}_{A_i})_{[k_i],J_2}\right)\right)^2 \det((\mathbf{H}_i)_{J_1}) \\ &\overset{(d)}{=} \det(\mathbf{Z}_{A_i}\mathbf{Z}_{A_i}^\top)\det((\mathbf{H}_i)_{J_1}), \end{aligned} \tag{20}$$

where (a) applies Cauchy–Schwarz inequality and (b)(c)(d) apply Cauchy–Binet formula. Substituting Eq. 19 in Eq. 18 completes the proof. ■

Now we consider $\hat{\mathbf{H}}_i$ denotes the approximate $\mathbf{H}_i$ by its sparse representation. To preserve the determinant of the upper bound, we should ensure

$$\sum_{J_1 \in \binom{[m]}{k_i}} \det((\hat{\mathbf{H}}_i)_{J_1}) \overset{(a)}{\approx} \sum_{J_1 \in \binom{[m]}{k_i}} \det((\mathbf{H}_i)_{J_1}). \tag{21}$$

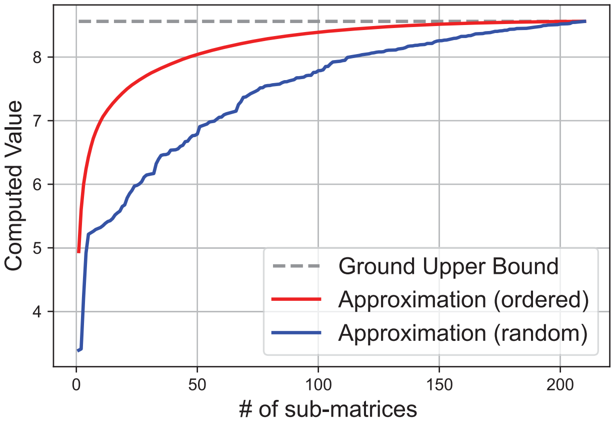


Figure 6. An exemplary visualization of approximation in Eq. 21. We generate $\mathbf{Z}_{A_i}, \mathbf{Z}_{Y_i} \in \mathbb{R}^{4\times 10}$. $\mathbf{H}_i$ is accordingly computed based on $\mathbf{Z}_{Y_i}$ by Eq. 16. The ground upper bound is computed by the left side of Eq. 21. Now, J_1 is a 4-th combination of $[10]$, resulting in a total of $10!/(4!(10-4)!) = 210$ sub-matrices. The red line presents a case where we select n_H sub-matrices with largest $\det((\hat{\mathbf{H}}_i)_{J_1})$ while the blue line denotes we randomly select n_H sub-matrices. It is observed that choosing the first n_H largest sub-matrices can obtain a lower error than randomly choosing these sub-matrices.

If a sub-matrix $(\mathbf{H}_i)_{\mathcal{C}}$ is allowed to transmit with a constraint $|\mathcal{C}| = r_0$, to minimize the difference between the left term and right term of (a) in Eq. 21, it is not feasible to traverse all possible sub-matrices when r_0 and k_i are a little bit large (i.e. $k_i!/(r_0!(k_i - r_0)!)$ combinations); however, we can immediately obtain a good sub-optimal solution by DPP greedy search as $\mathcal{C}^* = \text{MAP-DPP}(\mathbf{H}_i, r_0)$, since it has to be selected at least the first $r_0 - k_i + 1$ largest $\det((\mathbf{H}_i)_{J_1})$ in the greedy search. Here $\mathcal{C}^*$ denotes the index set of selected representative dimensions. Fig. 6 demonstrates the approximation by choosing n_H first largest sub-matrices $\det((\hat{\mathbf{H}}_i)_{J_1})$ can obtain a lower error than random choose these sub-matrices.

We define tolerable sparsity $R \times m$ for compression as the number of elements that can be losslessly transmitted from the center to each source. Transmitting the symmetric matrix $(\mathbf{H}_i)_{\mathcal{C}^*}$ requires $(r_0^2 + r_0)/2$ elements, which corresponds to the number of elements in the lower triangular matrix of $(\mathbf{H}_i)_{\mathcal{C}^*}$. Furthermore, in cases where additional sparsity can be utilized for compression purposes, we can compress the residual matrix $(\mathbf{H}_i)_{\bar{\mathcal{C}}^*}$ through singular value decomposition (SVD). Here $\bar{\mathcal{C}}^* = [m] \setminus \mathcal{C}^*$. By considering only the first r_1 singular vectors and values, we only require a sparsity of $r_1 m$ for the compression. Hence, the constraint on the tolerable sparsity R_m can be expressed as $(r_0^2 + r_0)/2 + r_1 m \leq Rm$. Mathematically,

$$\hat{\mathbf{H}}_i = (\mathbf{H}_i)_{\mathcal{C}^*} + \mathbf{V}(:, 1:r_1)\text{diag}(\lambda_1, \cdots, \lambda_{r_1})\mathbf{V}^\top(:, 1:r_1) \tag{22}$$

$$\begin{aligned} s.t. \quad & (r_0^2 + r_0)/2 + r_1 m \leq Rm, \\ & \mathcal{C}^* = \text{MAP-DPP}(\mathbf{H}_i, r_0), \\ & \mathbf{V}\text{diag}(\lambda_1, \cdots, \lambda_m)\mathbf{V}^\top = \text{svd}(\mathbf{H}_i - (\mathbf{H}_i)_{\mathcal{C}^*}). \end{aligned}$$

Therefore, the data samples in each source can be pre-coded as $\widetilde{\mathbf{Z}}_{S_i} = \mathbf{Z}_{S_i}\hat{\mathbf{H}}_i^{1/2}$. In fact, since the CSI information is not completely reliable, we precode the data samples conservatively and use a momentum way, which is presented as

$$\widetilde{\mathbf{Z}}_{S_i} = \mathbf{Z}_{S_i}\mathbf{W}_i = \mathbf{Z}_{S_i}(\mathbf{I} + \hat{\mathbf{H}}_i^{1/2}). \tag{23}$$

A summary of our approach is shown in Algorithm 1.

Algorithm 1 DDPP: MAP Inference for MAP for Distributed Data Source.

```
Input: Source data $\mathbf{Z}_{S_1}, \cdots, \mathbf{Z}_{S_N}$, Center information $\mathcal{Y} = \{Y_1, \cdots, Y_N\}$, the number of items selected in each interval $k_i$ for each source. Sparsity parameters $r_0, r_1$. The index set of selection $\mathcal{A}$.
Output: The updated index set of selection $\mathcal{A}$.
  for $i$ in $1:N$ do
    #In each source $i$. #All sources do these steps in parallel.
    Computed CSI information $\hat{\mathbf{H}}_i$ based on Eq. 22.
    Pre-code data in source $i$: $\widetilde{\mathbf{Z}}_{S_i} = \mathbf{Z}_{S_i}(\mathbf{I} + \hat{\mathbf{H}}_i^{1/2})$.
    #$\mathbf{Z}_{S_i}$ and $\widetilde{\mathbf{Z}}_{S_i}$ are the original and pre-coded data matrices, respectively.
    Compute new Gram matrix $\widetilde{\mathbf{L}}_{S_i}$: $\widetilde{\mathbf{L}}_{S_i} = \widetilde{\mathbf{Z}}_{S_i}\widetilde{\mathbf{Z}}_{S_i}^\top$.
    $A_i \leftarrow \emptyset$. #Init. index set for selection.
    while $|A_i| \leq k_i$ do
      #Selection based on DPP MAP. $j$ is the index of the selected sample.
      $j = \arg\max_{i \in S_i \setminus A_i} \log\det\left(\widetilde{\mathbf{L}}_{A_i \cup \{i\}}\right) - \log\det\left(\widetilde{\mathbf{L}}_{A_i}\right)$.
      if $j \notin \mathcal{A}$ then $A_i \leftarrow A_i \cup j$ end if
    end while
  end for
  $\mathcal{A} \leftarrow \mathcal{A} \cup A_i \cup \cdots \cup A_N$.
```

D. Complexity Analysis

In this section, we evaluate the computational complexity of the proposed method during one interval. We consider the added computational load in both sources and the center. In each source, the computation cost includes two terms for i) pre-coding via matrix multiplication (i.e. $\widetilde{\mathbf{Z}}_{S_i} = \mathbf{Z}_{S_i}(\mathbf{I} + \hat{\mathbf{H}}_i^{1/2})$), which is $\mathcal{O}(n_i m^2)$, and ii) generating candidates in each source by DPP, which is $\mathcal{O}(n_i^3) + \mathcal{O}(n_i k_i^2)$. Recalling that n_i and k_i denote the number of total samples and the number of selected samples in source i, we have $k_i < n_i$

and the overall computation complexity in each source can be approximated as follows $\mathcal{O}(n_i m^2) + \mathcal{O}(n_i^3) + \mathcal{O}(n_i k_i^2) = \mathcal{O}(n_i \max(m^2, n_i^2))$. All sources perform the computations independently and in parallel. If the total number of samples $n = n_1 + n_2 + \cdots n_N$ is fixed, when the number of sources N increases, we have $n_i = n/N \ll m$ and the complexity reduces to $\mathcal{O}(n_i m^2)$. In the center, the complexity comes from the computation of CSI (Eq. 22), which requires $\mathcal{O}(m^3) + \mathcal{O}(m r_0^2)$ to generate $\mathcal{C}^*$ and then a complexity of $\mathcal{O}(m^3)$ for SVD decomposition. Noting that $m \gg r_0$, the computational complexity in the center becomes $\mathcal{O}(m^3)$. In summary, the added computation load is $\mathcal{O}(n_i m^2)$ per source and $\mathcal{O}(m^3)$ for the center. Fortunately, the computation load is polynomial in the dimension of the data samples m and grows only linearly with the number of samples n_i. In most practical systems, the computational load of sources is more important because the central processing servers are typically equipped with higher computational resources. It is noteworthy that the computational cost per source in our method ($\mathcal{O}(n_i m^2)$) is in the same order as the other competitors, such as GreeDi [28]. with the additional benefit of our method in replacing heavy transmission of raw samples with lightweight diversity-representing messages.

VII. Experiment

A. Comparison Method

In our experiments, we use the exact greedy search proposed in [22] across all samples as the **Ground Truth**. We consider multiple alternative methods for comparison, including

- **GreeDi**: A two-round method, the most known distributed solution, in which, in the first round, each source greedily finds a set of size αk_T samples, and in the second round, it performs another greedy search on all candidates $N\alpha k_T$ from the previous round [28]. To meet a zero-communication overhead policy, we set $\alpha = 1/N$. It is also equivalent to our method without using the feedback mechanism.
- **MaxDiv Source**: It performs the exact greedy search in one source with the largest Information-theoretical diversity measured as $\log\det(\mathbf{I} + \frac{m}{|S_i|\epsilon}\mathbf{Z}_{S_i}^\top \mathbf{Z}_{S_i})$ [53].
- **Random Selection**: It involves random selection of samples by each source. While the total number of samples is the same as ours k_T, the number of samples selected by each source can be different.
- **Stratified Sampling**: Similar to random selection, but with an equal number of samples selected by each source (k_T/N).
- **Greedymax**: A two-round method, in the first round, each source greedily finds a set of size k_T samples, and in the second round, the set from the source with the maximum diversity, is sent to the center [28].

B. Dataset and Experiment Setup

The first experiment involves diversity evaluation using two datasets, CIFAR10 [54] and CIFAR100 [54]. Image datasets were preferred due to the following reasons: i) they have relatively high dimensions, which enables DPP to choose a subset with a larger number of samples, ii) semantic features can be easily obtained by pre-trained models, and iii) images are the predominant data type used in many practical applications, including our own projects of drone-based aerial monitoring [55], [56] and AI-based traffic monitoring [1], [42]. As a proof-of-concept experiment, we used a pre-trained ResNet-18[1] to extract the latent features of images and set the number of dimensions to $m = 512$. For the sake of completeness, we consider different numbers of sources $N = 5, 10, 12, 15, 20$. Each source includes a non-overlapping set of 500 different samples. We executed all of our experiments on a node of a cluster with an Intel(R) Xeon(R) Gold 6148 CPU with 125 gigabytes of memory. Note that the primary algorithm does not require a GPU to operate. For simplicity, we consider selecting a total of $k_T = 120$ samples in $t_T = 2$ intervals. Therefore, a total of 60 samples is selected in each interval. Note that the feedback mechanism is utilized only once per interval. We set the tolerable sparsity, defined in Section C, to $R = 0.75 \times k_T/t_T = 45$. Note that $R = k_T/t_T$ leads to a trivial scenario where we can send all the previously received samples back to each source.

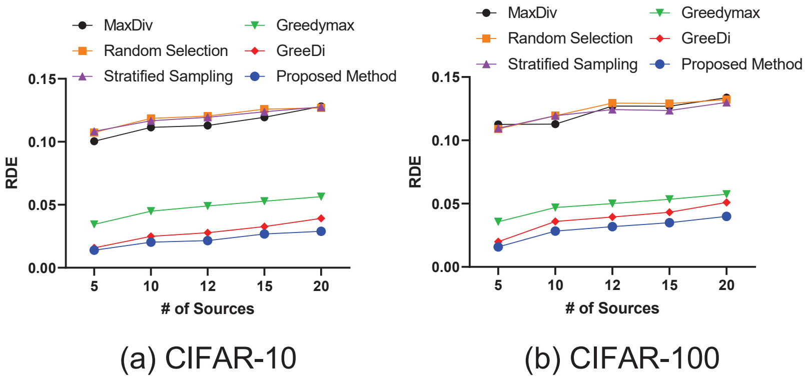
C. Comparative Results

The original DPP-diversity for a selected subset $\widetilde{A}$ is defined as $\det(\mathbf{Z}_{\widetilde{A}}\mathbf{Z}_{\widetilde{A}}^\top)$, where a higher value represents a higher level of diversity. For comparison convenience, the performance is presented as the **Relative Diversity Error (RDE)** with respect to the ground truth. Specifically, if the ground truth is A^* and the inference subset by one approach is $\widetilde{A}$. Then, the RDE is defined as

$$1 - \log\det(\mathbf{Z}_{A^*}\mathbf{Z}_{A^*}^\top)/\log\det(\mathbf{Z}_{\widetilde{A}}\mathbf{Z}_{\widetilde{A}}^\top), \qquad (24)$$

with a range of 0 to 1 (**the lower is better**). The results from running 20 times are shown in Table 3 and Fig. 7(a-b). The main observation is that our approach outperforms all baselines and alternative methods, exhibiting a considerable gain consistently for different numbers of sources on both datasets. Our approach significantly improves upon the random, stratified random and MaxDiv selection with a considerable reduction of 75% in RDE. More importantly, we found the feedback mechanism can decrease RDE varying from 12% to 26% in CIFAR-10 and 19% to 21% in CIFAR-100 compared to the method without feedback (i.e. GreeDi). The diversity increases for all methods by collecting samples from more sources as expected, but the performance superority for our method does not diminish. We also perform a two-sample t-Test for the selection performance to demonstrate the reliability of our method. According to the results shown in Table 4, the P-value is tiny and much smaller than 0.05, which indicates our method indeed has a statistical superiority over the second-best method, GreeDi.

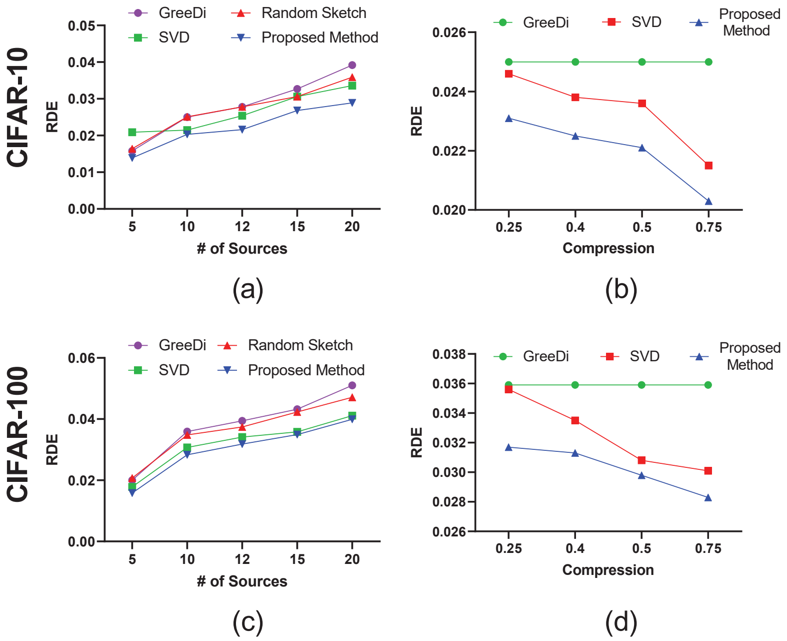
[1] This model can be found at https://pytorch.org/vision/stable/models.html.



(a) CIFAR-10

(b) CIFAR-100

Figure 7. Comparison of the performance (↓) of different selection strategies on CIFAR-10 and CIFAR-100 datasets in terms of Relative Diversity Error (RDE).



(a)

(b)

(c)

(d)

Figure 8. Comparison of the performance (↓) of different compression strategies (a)(c) and tolerable sparsity (b)(d) on CIFAR-10 and CIFAR-100 datasets.

D. Ablation Analysis

We conduct the following experiments to investigate the benefits of sensing the compressed version of CSI messages, $\mathbf{H_i}$, through the feedback channel. , we compare our method against the following alternative method,

- **SVD** replaces the proposed way of compressing $\mathbf{H_i}$ with an SVD-based compression (i.e. using the first R singular vectors of $\mathbf{H_i}$).
- **Random Sketch** generates set $\mathcal{C}$ by randomly sampling from the set $[m]$ in Eq. 22. As before, $[m] = \{1, \cdots, m\}$ represents the index set of dimensions.

The results are shown in Table 5 and Fig. 8(a)(c), confirming that both SVD and the proposed method can be used to improve the selection over Random Sketch method. Nonetheless, our method outperforms both two alternative compression strategies,SVD and Random Sketch. We found Random Sketch, at some time, is even worse than the selection without any feedback (GreeDi), indicating that a naive way of forming feedback messages can be misleading.

Additionally, we evaluate the impact of different tolerable sparsity (R) in Table 6 and Fig. 8(b)(d). The results show that our method can outperform SVD-based compression in different tolerable sparsity. Even at 0.25 level, our methods obtain improvement and achieve a reduction of 10% in RDE compared to GreeMi and SVD in CIFAR-100 dataset.

VIII. Potential Applications

In most practical data-driven AI-platforms, the ultimate goal is not only escalating the sampling diversity, but also to improve the users' Quality of Experience (QoE), reflected in the performance of the downstream learning-based tasks (e.g., classification, object detection, etc.). Translating diversity gain to learning quality often needs to adopt a proper

Table 3. The relative diversity error (RDE) of different selection methods. (↓): Lower is better.

Dataset	#ofCluster	5	10	12	15	20
CIFAR10	Ground Truth	0.0000	0.0000	0.0000	0.0000	0.0000
	MaxDiv Source	0.1004	0.1115	0.1130	0.1194	0.1281
	Random Selection	0.1074	0.1186	0.1205	0.1259	0.1270
	Stratified Sampling	0.1083	0.1167	0.1194	0.1238	0.1277
	Greedymax	0.0344	0.0450	0.0491	0.0528	0.0565
	GreeDi	0.0158	0.0250	0.0278	0.0327	0.0392
	DDPP (Proposed)	**0.0139**	**0.0203**	**0.0216**	**0.0268**	**0.0209**
CIFAR100	Ground Truth	0.0000	0.0000	0.0000	0.0000	0.0000
	MaxDiv Source	0.1126	0.1128	0.1271	0.1270	0.1337
	Random Selection	0.1090	0.1195	0.1295	0.1291	0.1322
	Stratified Sampling	0.1096	0.1195	0.1243	0.1236	0.1300
	Greedymax	0.0357	0.0470	0.0501	0.0534	0.0574
	GreeDi	0.0200	0.0359	0.0394	0.0432	0.0510
	DDPP (Proposed)	**0.0158**	**0.0283**	**0.0318**	**0.0349**	**0.0399**

Table 4. The P-value (↓) of two-sample t-Test between the proposed and GreeDI methods.

Dataset	#ofCluster	5	10	12	15	20
CIFAR10	GreeDI-DDPP (Proposed)	1.1E-04	1.1E-07	3.6E-09	6.3E-10	1.9E-16
CIFAR100	GreeDI-DDPP (Proposed)	1.6E-07	2.4E-11	1.4E-08	8.0E-08	9.1E-11

distance metric for DPP-based analysis. Designing a proper measure is out of the focus of this paper; however, as mentioned in Section B, we can simply use dot vectors among feature vectors, extracted by a pre-trained ResNet-18 as a low-dimensional semantic representation of data samples when calculating inter-sample correlations. This approach benefits both efficient selection and the resulting learning quality, while not requiring a custom-built similarity metric. It means that even a higher learning quality is attainable for each task if we use a tasks-specific distance metric and compromise the generalizability. This idea is commonly used to build *knowledge base* in the realm of semantic communication networks [44]. The following examples show the boosted learning performance when using our sample selection strategy.

Table 5. Comparison of the performance in RDE (↓) by different compression strategies with different numbers of sources.

Dataset	#ofCluster	5	10	12	15	20
CIFAR10	GreeDi	0.0158	0.0250	0.0278	0.0327	0.0392
	SVD	0.0289	0.0215	0.0254	0.0306	0.0336
	Random Sketch	0.0164	0.0251	0.0278	0.0306	0.0359
	DDPP (Proposed)	**0.0139**	**0.0203**	**0.0216**	**0.0268**	**0.0289**
CIFAR100	GreeDi	0.0200	0.0359	0.0394	0.0432	0.0510
	SVD	0.0178	0.0307	0.0341	0.0358	0.0411
	Random Sketch	0.0207	0.0348	0.0374	0.0423	0.0471
	DDPP (Proposed)	**0.0158**	**0.0283**	**0.0318**	**0.0349**	**0.0399**

Table 6. Comparison of the performance in RDE (↓) by different compression strategies with different tolerable sparsity.

Dataset	Sparsity ($\times k_T/t_T$)	0.25	0.4	0.5	0.75
CIFAR10	GreeDi	0.0250	0.0250	0.0250	0.0250
	SVD	0.0246	0.0238	0.0236	0.0215
	DDPP (Proposed)	**0.0231**	**0.0225**	**0.0221**	**0.0203**
CIFAR100	GreeDi	0.0359	0.0359	0.0359	0.0359
	SVD	0.0356	0.0335	0.0308	0.0301
	DDPP (Proposed)	**0.0317**	**0.0313**	**0.0298**	**0.0283**

A. Classification

This test involves conducting classification on CIFAR-10 and CIFAR-100 datasets. We use the k-nearest neighbors (KNN), a non-parametric method, to evaluate the representation of selected samples. The classification results on both datasets are shown in Tables 7 and 8, respectively. The results suggest that training the classifier with data samples selected by our method outperforms all other methods on both datasets. For example, our method achieves at least 3% higher accuracy on CIFAR-10. The F1 score improvement is at least 0.02. Fig. 9 presents the selected samples by different methods in CIFAR-10 visualized by Principal Component Analysis (PCA). It can be observed that our method, overall, selects a more diverse set of samples (shown by red circles), compared to random selection. This is the ground for achieving higher learning quality.

For the sake of completeness, we also investigate the classification results on Tiny-ImageNet [57], a very challenging large-scale dataset. For example, training a ResNet classifier on the entire dataset can only achieve about top-1 classification accuracy of 55%. In this experiment, we use pre-trained ResNet-50 to extract the latent feature and select 1200 samples in total. We compare our methods with the recent state-of-the-art data selection methods in the distributed setting. These methods include K-Center selection [58], submodular mutual information selection (MIV2) [59], density-aware selection (DACS) [60], and coverage-centric Selection (CCS) [61]. The results are shown in Table 9. Our methods consistently demonstrate superiority over other methods with an average improvement of 1.6% in accuracy and 2.5% improvement in F1 score across different numbers of sources. Additionally, our method obtains a substantial gain over the random selection with a 4%-6% improvement in accuracy. Another interesting observation is that although the recent selection methods (i.e., DACS and CCS) occasionally outperform GreeDi (our method without utilizing the feedback mechanism), our method, by leveraging the feedback mechanism, can easily compete with them. This further underscores the crucial role of the proposed feedback mechanism under distributed sources.

B. Traffic Sign Detection

In order to investigate a more practical scenario, we consider traffic sign detection in the realm of smart transportation [62]. It involves the identification and localization of traffic signs in images, an integral part of Autonomous Vehicles' (AVs) control stack, Advanced Driver Assistance Systems (ADAS), or traffic management enterprise. As a proof-of-concept, we use a small dataset [63] in our experiments. This dataset contains 877 images of 4 distinct classes, including Traffic Lights, Stops, Speed limit, and Crosswalk. We use 177 images

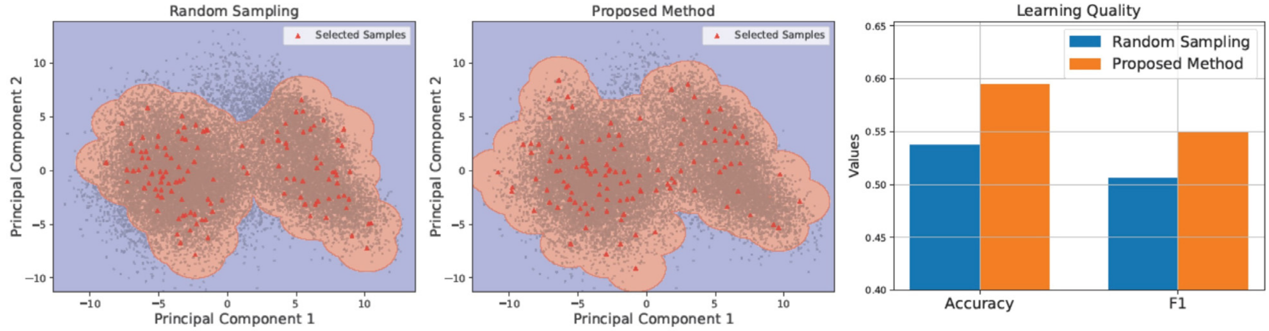


Figure 9. Selected data points by our method in comparison with random selection. We use the first two principal components of data for visualization. The selected samples are denoted by ▲ markers. We draw the $\varepsilon-$ball (i.e. $Ball(x_0, \varepsilon) = \{x_i : \|x_i - x_0\|_2 < \varepsilon\}$) for each selected point as shown in the red area to denote its coverage in the feature space. This concept is employed from Covering Number [20]. The proposed method presents better coverage, reflected in the elevated accuracy of a classifier that uses our method to select training samples.

Table 7. Comparison of classification results on CIFAR-10 by different strategies.

Method	#of Sources	5	10	12	15	20
Random	Accuracy (↑)	53.16	52.79	52.66	52.92	53.96
	F1 (↑)	0.533	0.483	0.499	0.500	0.519
Stratified Sampling	Accuracy (↑)	53.85	53.16	53.43	54.26	53.76
	F1 (↑)	0.517	0.512	0.509	0.519	0.514
GreeDi	Accuracy (↑)	55.52	54.84	54.62	54.65	53.05
	F1 (↑)	0.523	0.512	0.502	0.514	0.500
DDPP (Proposed)	Accuracy (↑)	**58.51**	**57.27**	**57.40**	**58.05**	**56.52**
	F1 (↑)	**0.560**	**0.540**	**0.543**	**0.554**	**0.538**

Table 8. Comparison of classification results on CIFAR-100 by different strategies.

Method	#of Sources	5	10	12	15	20
Random	Accuracy (↑)	46.71	45.58	44.26	44.15	42.43
	F1 (↑)	0.439	0.431	0.420	0.428	0.392
Stratified Sampling	Accuracy (↑)	46.52	46.05	43.36	44.33	44.55
	F1 (↑)	0.442	0.432	0.406	0.410	0.418
GreeDi	Accuracy (↑)	48.08	48.48	45.72	45.73	44.79
	F1 (↑)	0.441	0.456	0.414	0.399	0.391
DDPP (Proposed)	Accuracy (↑)	**49.68**	**49.02**	**47.22**	**47.78**	**46.29**
	F1 (↑)	**0.471**	**0.459**	**0.442**	**0.452**	**0.421**

as the test set. We set the number of sources $N = 10$, and each source contains 70 images. We select $k_T = 120$ samples from this dataset by using different selection strategies. Then, use the selected samples to train a state-of-the-art object detector YOLOv8 [64] from scratch. The performance of detection is evaluated by mean Average Precision (mAP) at different Intersection over Union (IoU) thresholds (0.5 and 0.5-0.95) and F1 score. The results are obtained by averaging over 10 runs. As shown in Table 10, increasing the diversity generally can enhance the detection performance. For example, with a 0.05 RDE decrease compared with Random selection, our method can obtain around 0.05 mAP improvement and 0.04 F1 score improvement, respectively. Fig. 10 demonstrates the selected images inference by the model trained on different selected data, which brings considerable improvement in both accuracy and confidence of the detection.

Table 9. Comparison of classification results on Tiny-ImageNet by different strategies.

Method	# of Sources	5	10	12	15	20
Random	Accuracy (↑)	27.66	28.01	28.06	27.37	28.01
	F1 (↑)	0.248	0.251	0.251	0.248	0.252
Stratified Sampling	Accuracy (↑)	28.14	27.88	28.40	28.45	27.78
	F1 (↑)	0.254	0.248	0.254	0.255	0.250
K-center	Accuracy (↑)	30.15	28.80	28.86	28.06	27.55
	F1 (↑)	0.292	0.285	0.280	0.273	0.267
MIV2	Accuracy (↑)	29.91	28.43	28.55	28.28	28.30
	F1 (↑)	0.268	0.255	0.257	0.255	0.254
DACS	Accuracy (↑)	33.77	32.32	32.51	31.59	31.77
	F1 (↑)	0.285	0.270	0.269	0.262	0.270
CCS	Accuracy (↑)	33.15	32.60	32.66	31.57	31.44
	F1 (↑)	0.282	0.273	0.271	0.265	0.263
GreeDi	Accuracy (↑)	32.36	32.66	32.52	31.82	31.86
	F1 (↑)	0.280	0.274	0.271	0.266	0.266
DDPP (Proposed)	Accuracy (↑)	**34.66**	**34.61**	**33.66**	**33.40**	**32.92**
	F1 (↑)	**0.305**	**0.302**	**0.295**	**0.292**	**0.287**

Table 10. Comparison of performance on object detection of different selection strategies.

Sampling Method		Random	Stratified Sampling	GreeMi	DDPP (Proposed)
	RDE (↓)	0.1077	0.0986	0.0795	**0.0535**
Detection	mAP@50 (↑)	0.5920	0.5923	0.6364	**0.6488**
	mAP@50-95 (↑)	0.4427	0.4553	0.4734	**0.4961**
	F1 score (↑)	0.5724	0.5811	0.6016	**0.6120**

C. Car Crash Detection via Multiple-Instance Learning

To show the utility of our method in the more complex problem of video analysis, we develop an experiment for video-based crash detection. In real surveillance videos, the task of classifying whether a crash has occurred is of paramount importance, but what is even more critical is the rapid detection of the precise moment when a crash occurs. However, this presents a significant challenge as annotating each individual frame in the video is a time-consuming

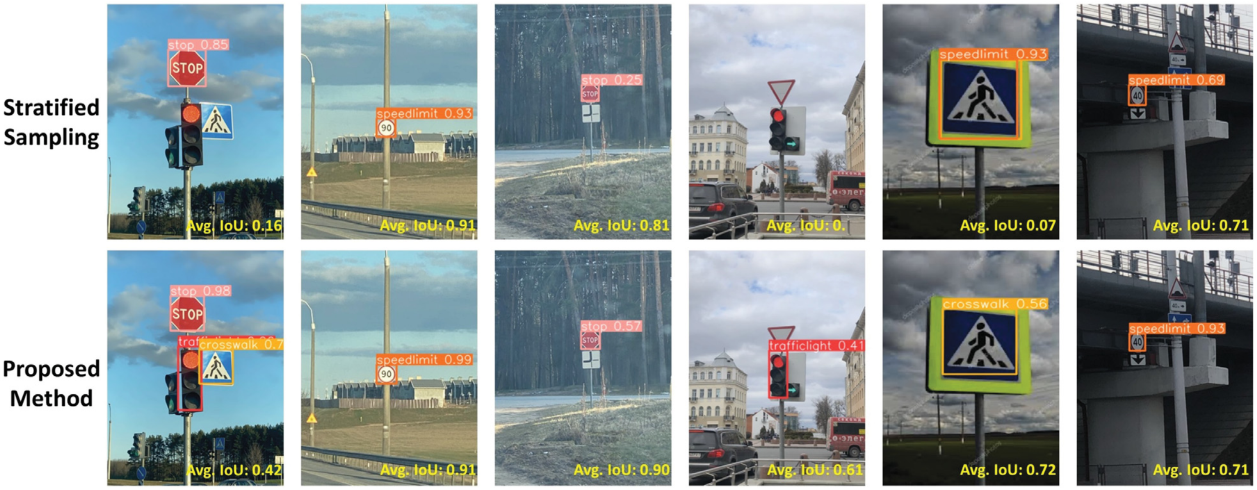


Figure 10. The application of our method in traffic sign interpretation. The selected results of detection by training a YOLOv8 detector from scratch using data selected by Top: *Stratified Sampling* and Bottom: Our proposed method. Our method demonstrates higher confidence and higher accuracy in the detection.

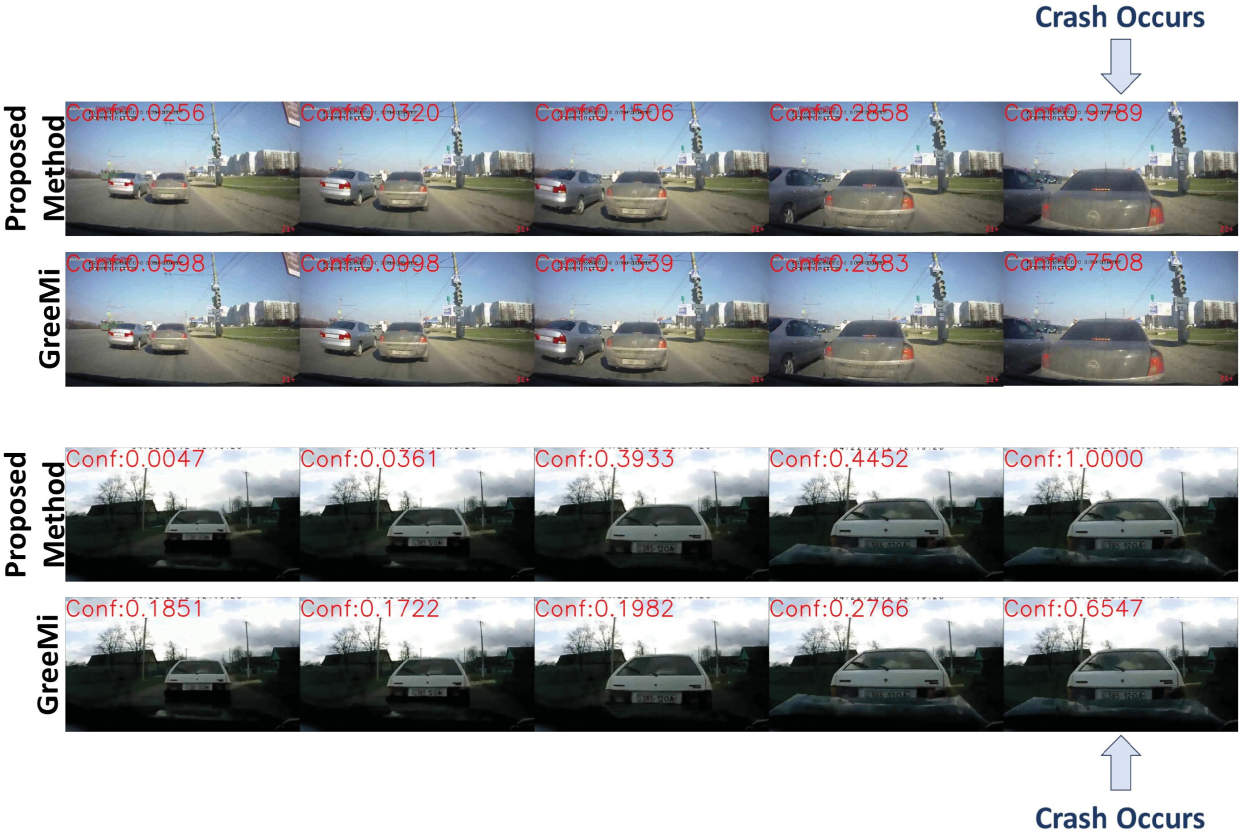


Figure 11. Video-based crash detection. Training videos are selected based on their frame diversity with respect to other videos using different methods. The confidence of the crash occurs provided by the attention map in sequential frames by two selection methods.

Table 11. The learning quality of multiple-instance learning by different data selection strategies.

Sampling Method		Random	Stratified Sampling	GreeMi	DDPP (Proposed)
RDE (↓)		0.1495	0.1467	0.0960	0.0939
MIL	Accuracy (↑)	0.8251	0.8247	0.8359	**0.8548**
	F1 score (↑)	0.8295	0.8321	0.8385	**0.8549**
	AUC (↑)	0.9066	0.9038	0.9094	**0.9228**

and labor-intensive task, made even more complex by the varying lengths of different videos. To address this issue, a practical solution is to leverage video-level annotation and employ Multiple-Instance Learning (MIL) [65], which is a weakly-supervised method and widely used in Anomaly Detection [66]–[68].

In this problem, each frame in a surveillance video is treated as an instance within a bag and is assigned a binary label, either 0 (for negative) or 1 (for positive), based on the presence or absence of a crash event. The video, on the other hand, serves as a bag containing multiple instances, which are the frames. A key characteristic is that a bag is considered negative (labeled as 0) only if all its instances, meaning all the frames within the video, are negative. Otherwise, if any frame within the video is positive, the entire bag is labeled as positive (1), indicating the presence of a crash event. In this work, an attention-based MIL method [65] is utilized. This method likely involves the use of attention mechanisms to weigh the importance of different instances (frames) within a bag (video) when making the final classification decision, effectively allowing the model to focus on the most informative frames for crash detection.

We use the Car Crash Dataset (CCD) [69], which contains traffic accident videos captured by dashcams mounted on driving vehicles. We choose $N = 10$ and each source contains 100 videos. We select $k_T = 80$ videos using $t_T = 2$ intervals, using various selection methods. The diversity and learning quality of selected samples are shown in Table 11, which demonstrates an improved accuracy, F1 score, and AUC for our selection method. The frame-level prediction confidence of the crash obtained by the attention map from the trained model is shown in Fig. 11, which demonstrates in both exemplary frame sequences, the model trained by samples selected by our method has the more accurate localization of the crash occurs.

IX. Conclusion

DPP is a formal method to enhance data diversity for learning-based systems. However, it requires access to the entire dataset in one place, which limits its applicability to diverse sources in real-world applications. To address this key challenge, we implemented a DPP MAP inference for distributed data from multiple sources as a universal diversity-maximizing data-sharing strategy for distributed sources that only requires a lightweight feedback channel from the center to the sources with no cross-source communication requirement. To this end, a novel scheduling policy, inspired by MIMO systems, is proposed. Specifically, we demonstrated that the lower bound of the original diversity maximization problem that maximizes global diversity can be decomposed into a sum of factors that enables distributed selection. Additionally, approximating the lower bound to the original problem can be treated as receiving *CSI* and *pre-coding*. Under communication bandwidth constraints, we derive a sparse CSI representation to preserve the determinant via the Cauchy–Binet formula. Our experiments demonstrate that our scalable approach can compete with all alternative methods in various datasets. Moreover, as a proof-of-concept, we show that with proper distance measures, pursuing diversity can translate into improving learning quality in multiple applications, including multi-level classification, object detection, and multiple-instance learning. We expect our approach can substantially influence the design of future AI-based networking platforms, which require efficient handling of massive datasets split across distributed sources.

References

[1] A. Sarlak, A. Razi, X. Chen, and R. Amin, "Diversity maximized scheduling in roadside units for traffic monitoring applications," in *2023 IEEE 48th Conference on Local Computer Networks (LCN)*. IEEE, 2023, pp. 1–4.

[2] J. Garau Guzman and V. M. Baeza, "Enhancing urban mobility through traffic management with uavs and vlc technologies," *Drones*, vol. 8, no. 1, p. 7, 2023.

[3] A. B. Haque, B. Bhushan, and G. Dhiman, "Conceptualizing smart city applications: Requirements, architecture, security issues, and emerging trends," *Expert Systems*, vol. 39, no. 5, p. e12753, 2022.

[4] M. Chen, W. Li, Y. Hao, Y. Qian, and I. Humar, "Edge cognitive computing based smart healthcare system," *Future Generation Computer Systems*, vol. 86, pp. 403–411, 2018.

[5] S. Tanwar, K. Parekh, and R. Evans, "Blockchain-based electronic healthcare record system for healthcare 4.0 applications," *Journal of Information Security and Applications*, vol. 50, p. 102407, 2020.

[6] M. A. Rahman, M. S. Hossain, A. J. Showail, N. A. Alrajeh, and A. Ghoneim, "Ai-enabled iiot for live smart city event monitoring," *IEEE Internet of Things Journal*, 2021.

[7] D. Valle-Cruz, V. Fernandez-Cortez, and J. R. Gil-Garcia, "From e-budgeting to smart budgeting: Exploring the potential of artificial intelligence in government decision-making for resource allocation," *Government Information Quarterly*, vol. 39, no. 2, p. 101644, 2022.

[8] Z. Yang, L. Feng, F. Zhou, X. Qiu, and W. Li, "Ergodic capacity analysis of irs aided wireless-powered relay communication network," *IEEE Transactions on Cognitive Communications and Networking*, 2023.

[9] N. Van Huynh, D. T. Hoang, X. Lu, D. Niyato, P. Wang, and D. I. Kim, "Ambient backscatter communications: A contemporary survey," *IEEE Communications surveys & tutorials*, vol. 20, no. 4, pp. 2889–2922, 2018.

[10] H. Yang, X. Yuan, J. Fang, and Y.-C. Liang, "Reconfigurable intelligent surface aided constant-envelope wireless power transfer," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1347–1361, 2021.

[11] K. Sayood, *Introduction to data compression*. Morgan Kaufmann, 2017.

[12] X. Luo, H.-H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," *IEEE Wireless Communications*, vol. 29, no. 1, pp. 210–219, 2022.

[13] H. G. Abreha, M. Hayajneh, and M. A. Serhani, "Federated learning in edge computing: a systematic survey," *Sensors*, vol. 22, no. 2, p. 450, 2022.

[14] G. Lan, X.-Y. Liu, Y. Zhang, and X. Wang, "Communication-efficient federated learning for resource-constrained edge devices," *IEEE Transactions on Machine Learning in Communications and Networking*, 2023.

[15] Y. Yang, L. Feng, Y. Sun, Y. Li, W. Li, and M. A. Imran, "Multi-cluster cooperative offloading for vr task: A marl approach with graph embedding," *IEEE Transactions on Mobile Computing*, 2024.

[16] B. O. Ayinde, T. Inanc, and J. M. Zurada, "Regularizing deep neural networks by enhancing diversity in feature extraction," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2650–2661, 2019.

[17] X. Wu, S. Zhang, Q. Zhou, Z. Yang, C. Zhao, and L. J. Latecki, "Entropy minimization versus diversity maximization for domain adaptation," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[18] Y. Yu, S. Khadivi, and J. Xu, "Can data diversity enhance learning generalization?" in *Proceedings of the 29th international conference on computational linguistics*, 2022, pp. 4933–4945.

[19] S. Zhou, J. Wang, L. Wang, X. Wan, S. Hui, and N. Zheng, "Inverse adversarial diversity learning for network ensemble," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[20] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[21] A. Kulesza, B. Taskar *et al.*, "Determinantal point processes for machine learning," *Foundations and Trends® in Machine Learning*, vol. 5, no. 2–3, pp. 123–286, 2012.

[22] L. Chen, G. Zhang, and E. Zhou, "Fast greedy map inference for determinantal point process to improve recommendation diversity," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[23] M. Derezinski, D. Calandriello, and M. Valko, "Exact sampling of determinantal point processes with sublinear time preprocessing," *Advances in neural information processing systems*, vol. 32, 2019.

[24] D. Calandriello, M. Derezinski, and M. Valko, "Sampling from a k-dpp without looking at all items," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6889–6899, 2020.

[25] M. Dereziński, "Fast determinantal point processes via distortion-free intermediate sampling," in *Conference on Learning Theory*. PMLR, 2019, pp. 1029–1049.

[26] I. Han and J. Gillenwater, "Map inference for customized determinantal point processes via maximum inner product search," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2797–2807.

[27] A. Bhaskara, A. Karbasi, S. Lattanzi, and M. Zadimoghaddam, "Online map inference of determinantal point processes," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3419–3429, 2020.

[28] B. Mirzasoleiman, A. Karbasi, R. Sarkar, and A. Krause, "Distributed submodular maximization," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 8330–8373, 2016.

[29] M. Vu and A. Paulraj, "Mimo wireless linear precoding," *IEEE Signal Processing Magazine*, vol. 24, no. 5, pp. 86–105, 2007.

[30] N. Fatema, G. Hua, Y. Xiang, D. Peng, and I. Natgunanathan, "Massive mimo linear precoding: A survey," *IEEE systems journal*, vol. 12, no. 4, pp. 3920–3931, 2017.

[31] V. M. Baeza and A. G. Armada, *Noncoherent Massive MIMO*. John Wiley & Sons, Ltd, 2020, pp. 1–28. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119471509.w5GRef027

[32] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (discus): Design and construction," *IEEE transactions on information theory*, vol. 49, no. 3, pp. 626–643, 2003.

[33] D. B. Kurka and D. Gündüz, "Deepjscc-f: Deep joint source-channel coding of images with feedback," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 178–193, 2020.

[34] Z. Gao, L. Dai, S. Han, I. Chih-Lin, Z. Wang, and L. Hanzo, "Compressive sensing techniques for next-generation wireless communications," *IEEE Wireless Communications*, vol. 25, no. 3, pp. 144–153, 2018.

[35] H. M. Nguyen and G. H. Glover, "A modified generalized series approach: application to sparsely sampled fmri," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2867–2877, 2013.

[36] X. Guo, Y. Lu, F. Wu, W. Gao, and S. Li, "Distributed multi-view video coding," in *Visual Communications and Image Processing 2006*, vol. 6077. SPIE, 2006, pp. 290–297.

[37] N. Gehrig and P. L. Dragotti, "Distributed compression of multi-view images using a geometrical coding approach," in *2007 IEEE International Conference on Image Processing*, vol. 6. IEEE, 2007, pp. VI–421.

[38] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang, and S. Wang, "Image and video compression with neural networks: A review," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1683–1698, 2019.

[39] O. Rippel, S. Nair, C. Lew, S. Branson, A. G. Anderson, and L. Bourdev, "Learned video compression," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3454–3463.

[40] T.-Y. Tung and D. Gündüz, "Deepwive: Deep-learning-aided wireless video transmission," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 9, pp. 2570–2583, 2022.

[41] A. Antsiferova, S. Lavrushkin, M. Smirnov, A. Gushchin, D. Vatolin, and D. Kulikov, "Video compression dataset and benchmark of learning-based video-quality metrics," *Advances in Neural Information Processing Systems*, vol. 35, pp. 13 814–13 825, 2022.

[42] X. Chen, H. Wang, A. Razi, B. Russo, J. Pacheco, J. Roberts, J. Wishart, L. Head, and A. G. Baca, "Network-level safety metrics for overall traffic safety assessment: A case study," *IEEE Access*, 2022.

[43] L. Xia, Y. Sun, D. Niyato, D. Feng, L. Feng, and M. A. Imran, "xurllc-aware service provisioning in vehicular networks: A semantic communication perspective," *IEEE Transactions on Wireless Communications*, 2023.

[44] G. Shi, Y. Xiao, Y. Li, and X. Xie, "From semantic communication to semantic-aware networking: Model, architecture, and open problems," *IEEE Communications Magazine*, vol. 59, no. 8, pp. 44–50, 2021.

[45] V. Hassija, V. Chawla, V. Chamola, and B. Sikdar, "Incentivization and aggregation schemes for federated learning applications," *IEEE Transactions on Machine Learning in Communications and Networking*, 2023.

[46] M. Mestoukirdi, M. Zecchin, D. Gesbert, and Q. Li, "User-centric federated learning: Trading off wireless resources for personalization," *IEEE Transactions on Machine Learning in Communications and Networking*, vol. 1, pp. 346–359, 2023.

[47] J. Xiao, X. Tang, and S. Lu, "Privacy-preserving federated class-incremental learning," *IEEE Transactions on Machine Learning in Communications and Networking*, 2023.

[48] L. Feng, Y. Zhao, S. Guo, X. Qiu, W. Li, and P. Yu, "Bafl: A blockchain-based asynchronous federated learning framework," *IEEE Transactions on Computers*, vol. 71, no. 5, pp. 1092–1103, 2021.

[49] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, "In-edge ai: Intelligentizing mobile edge computing, caching and communication by federated learning," *Ieee Network*, vol. 33, no. 5, pp. 156–165, 2019.

[50] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.

[51] D. Gesbert, M. Shafi, D.-s. Shiu, P. J. Smith, and A. Naguib, "From theory to practice: An overview of mimo space-time coded wireless systems," *IEEE Journal on selected areas in Communications*, vol. 21, no. 3, pp. 281–302, 2003.

[52] K. B. Petersen, M. S. Pedersen *et al.*, "The matrix cookbook," *Technical University of Denmark*, vol. 7, no. 15, p. 510, 2008.

[53] A. Y. Yang, J. Wright, Y. Ma, and S. S. Sastry, "Unsupervised segmentation of natural images via lossy data compression," *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 212–225, 2008.

[54] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[55] S. P. H. Boroujeni and A. Razi, "Ic-gan: An improved conditional generative adversarial network for rgb-to-ir image translation with applications to forest fire monitoring," *Expert Systems with Applications*, p. 121962, 2023.

[56] S. P. H. Boroujeni, A. Razi, S. Khoshdel, F. Afghah, J. L. Coen, L. O'Neill, P. Fule, A. Watts, N.-M. T. Kokolakis, and K. G. Vamvoudakis, "A comprehensive survey of research towards ai-enabled unmanned aerial systems in pre-, active-, and post-wildfire management," *Information Fusion*, p. 102369, 2024.

[57] Y. Le and X. Yang, "Tiny imagenet visual recognition challenge," *CS 231N*, vol. 7, no. 7, p. 3, 2015.

[58] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=H1aIuk-RW

[59] S. Kothawade, V. Kaushal, G. Ramakrishnan, J. Bilmes, and R. Iyer, "Submodular mutual information for targeted data subset selection," *ICLR 2021 From Shallow to Deep: Overcoming Limited and Adverse Data Workshop*, 2021.

[60] Y. Kim and B. Shin, "In defense of core-set: A density-aware core-set selection for active learning," in *Proceedings of the 28th ACM*

SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 804–812.

[61] H. Zheng, R. Liu, F. Lai, and A. Prakash, “Coverage-centric coreset selection for high pruning rates,” in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=QwKvL6wC8Yi

[62] A. Razi, X. Chen, H. Li, H. Wang, B. Russo, Y. Chen, and H. Yu, “Deep learning serves traffic safety analysis: A forward-looking review,” *IET Intelligent Transport Systems*, vol. 17, no. 1, pp. 22–71, 2023.

[63] “Road signs dataset.” [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/road-sign-detection

[64] G. Jocher, A. Chaurasia, and J. Qiu, “YOLO by Ultralytics,” Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[65] M. Ilse, J. Tomczak, and M. Welling, “Attention-based deep multiple instance learning,” in *International conference on machine learning*. PMLR, 2018, pp. 2127–2136.

[66] W. Zhu, P. Qiu, N. Lepore, O. M. Dumitrascu, and Y. Wang, “Self-supervised equivariant regularization reconciles multiple-instance learning: Joint referable diabetic retinopathy classification and lesion segmentation,” in *18th International Symposium on Medical Information Processing and Analysis*, vol. 12567. SPIE, 2023, pp. 100–107.

[67] Y. Gong, C. Wang, X. Dai, S. Yu, L. Xiang, and J. Wu, “Multi-scale continuity-aware refinement network for weakly supervised video anomaly detection,” in *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2022, pp. 1–6.

[68] H. Lv, Z. Yue, Q. Sun, B. Luo, Z. Cui, and H. Zhang, “Unbiased multiple instance learning for weakly supervised video anomaly detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8022–8031.

[69] W. Bao, Q. Yu, and Y. Kong, “Uncertainty-based traffic accident anticipation with spatio-temporal relational learning,” in *ACM Multimedia Conference*, May 2020.