Progressive Adversarial Multi-View Alignment for Unsupervised Embedded Feature Selection with Linear Complexity

Shixuan Zhou[1], **Yi Xiang**[1*], **Haoxiang Qin**[1], **Haining Wang**[1], **Yukuan Ma**[2] and **Han Huang**[3]
[1]South China University of Technology
[2]Airace Technology Co.,Ltd.
[3]Sun Yat-sen University
shixuanzhou0601@163.com, xiangyi@scut.edu.cn, klaetteqin@gmail.com, hining0203@163.com, martinup@126.com, hanhuang@foxmail.com

Abstract

Standard unsupervised multi-view feature selection (UMFS) methods for large datasets exhibit limitations in modeling the competition between cross-view alignment and intra-view diversity, resulting in suboptimal solutions and expensive computational costs. This challenge is exacerbated by the diverse signals inherent in complex samples, which tend to mask shared patterns, thus complicating the pursuit of an optimal trade-off. In this work, we propose a Progressive Adversarial Feature Selection (ProAd-FS) framework for large-scale multi-view learning, which formulates this static trade-off objective as a dynamic competitive process. To be specific, ProAd-FS develops an adversarial decoupled architecture that assigns these competing objectives to distinct inter-view and intra-view games, guided by a robust gated curriculum to prioritize the learning of underlying structures. The entire framework exhibits linear computational complexity in both sample size and feature dimension, and embeds structural sparse regularization for end-to-end optimization. Extensive experimental results show that ProAd-FS achieves state-of-the-art performance while being highly scalable as well, providing an effective solution for large-scale UMFS tasks.

1 Introduction

As object collections continue to expand, modern datasets increasingly involve objects being assigned multiple distinct feature sets, forming multi-view representations [Xu *et al.*, 2024; Khalafaoui *et al.*, 2025]. Such views typically appear as images paired with text descriptions [Feng *et al.*, 2024] or as sensor data drawn from different modalities [Fadadu *et al.*, 2022]. By providing complementary information, they enable a more comprehensive and profound understanding of underlying data structures, making multi-view representation learning a key research focus in computer vision, bioinformatics, and natural language processing [Li *et al.*, 2018; Fang *et al.*, 2023b]. However, the high dimensionality and inherent redundancy nature of raw multi-view data lead to high costs and challenges in fitting. Unsupervised multi-view feature selection (UMFS) addresses these challenges by selecting a compact and informative feature subset from all available views to enhance the performance and efficiency of downstream tasks [Dong *et al.*, 2025].

The fundamental challenges in UMFS involve achieving cross-view consistency while preserving intra-view diversity [Zhang *et al.*, 2024a; Liu *et al.*, 2025]. While maintaining consistency facilitates learning unified representations from cross-view knowledge, this process should not come at the expense of diversity, which constitutes the core discriminative value of individual viewpoints. Current approaches to cross-view issues predominantly rely on a single static objective to resolve this conflict, yielding unsatisfactory outcomes. For instance, canonical correlation analysis maximizes statistical correlations between views while compromising discriminative information [Xu *et al.*, 2019b; Xu *et al.*, 2019a], whereas joint matrix factorization reconstructs single-view spaces without considering inter-view spatial coherence, thereby undermining consistency [Hsieh *et al.*, 2019; Cao and Xie, 2024]. This trade-off between static objectives prevents models from effectively addressing both consistency and diversity challenges, thereby limiting their practical applications.

Another challenge stems from the massive scale of modern datasets, which creates a disconnect between theoretical algorithm design and practical implementation. Standard methods inherently involve super-linear complexity computations, facing scalability issues in both memory allocation and runtime. These limitations are particularly evident in fundamental algorithms like constructing pairwise similarity graphs, where quadratic expansion in computational complexity poses significant challenges for handling large datasets [Wang *et al.*, 2021]. Furthermore, it is also manifested in algorithms requiring iterative updates of complete data matrices, which demand computationally intensive matrix factorization [Zhou *et al.*, 2023a; Wu *et al.*, 2024]. Consequently, reliance on complex computations leads to nonlinear growth, creating a mismatch between theoretical algorithmic design and scalable practice.

This study proposes a Progressive Adversarial Feature Selection (ProAd-FS) framework to address computational and structural challenges. To be specific, ProAd-FS utilizes an

*Corresponding author.

adversarial framework to assign competing objectives into a new decoupled architecture, which consists of an inter-view game to learn view-invariant representations and an intra-view game to retain feature diversity. We further develop a Robust Gated Curriculum (RGC) to impose an easy-to-hard learning trajectory throughout the training process. This enables ProAd-FS to construct a robust shared baseline from basic data, progressively expanding through complex sample details, yielding stronger comprehensive discrimination. The entire framework is optimized via an end-to-end objective function embedding structured sparsity regularization through $\ell_{2,1}$-norm, demonstrating linear complexity for scalability across large datasets. Experimental results show ProAd-FS outperforms baseline methods on large-scale benchmark datasets of up to 70,000 samples, while running over more than an order of magnitude faster than baselines. In summary, our contributions are as follows.

1. We present a novel UMFS framework that leverages the adversarial game to formulate the static trade-off between cross-view consistency and intra-view diversity as a dynamic competitive process.
2. We design a gated curriculum, which imposes a hierarchical feature space discovery process to progressively guide the model towards a robust subset of features.
3. ProAd-FS demonstrates linear computational complexity in both sample size and feature dimensionality, with high scalability.

2 Related Works

Unsupervised Multi-view Feature Selection. Numerous UMFS methods enforce cross-view consistency by learning shared latent representations or pseudo-label spaces. Early studies used pseudo-class labels as a bridge between views [Tang *et al.*, 2013] or decomposed the label space into consensus and view-specific components [Tang *et al.*, 2021]. Other prominent studies employ adaptive learning to derive a unified similarity matrix that reflects cross-view commonalities. Building upon this paradigm, numerous methods focus on learning a common similarity matrix with view-specific weights [Hou *et al.*, 2017; Wan *et al.*, 2020], integrating latent adjacency representations [Zhou *et al.*, 2023b], or employing non-linear kernels and hashing to generate weak supervision [Hu *et al.*, 2025]. In terms of scalability, [Zhang *et al.*, 2024b] constructed bipartite graphs to reduce complexity to near-linear. Our work also achieves linear complexity, but unlike previous studies, ProAd-FS innovatively modifies the objective function by introducing a dynamic adversarial game mechanism, enabling a more scalable and robust feature selection process, especially in complex data environments.

Adversarial Representation Learning. Generative Adversarial Networks (GANs), proposed by [Goodfellow *et al.*, 2014], represent a powerful framework for representation learning. Existing applications are diverse, including learning rich visual representations with bidirectional GANs [Donahue and Simonyan, 2019], decoupling through mutual information maximization [Chen *et al.*, 2016], generating challenging training samples to learn robust encoders [Pan *et al.*, 2021], and ensuring discriminative representation in hybrid architectures [Zuo *et al.*, 2023]. While these methods typically rely on a single game to achieve their goals, ProAd-FS introduces an adversarial decoupled framework. By assigning consistency and diversity as distinct games to two independent processes, it transforms the static trade-off in UMFS into dynamic competition.

Curriculum and Self-Paced Learning. The Curriculum Learning (CL) strategy proposed by [Bengio *et al.*, 2009] for progressively presenting training samples from simple to complex has been proven effective in improving model robustness. Self-Paced Learning (SPL) automates this training by dynamically selecting samples based on current loss values [Kumar *et al.*, 2010]. Other improvements to SPL have focused on encouraging sample diversity [Jiang *et al.*, 2014], establishing self-adaptive difficulty ratings [Kong *et al.*, 2021], mitigating outliers in feature selection [Li *et al.*, 2022], and ensuring high-difficulty training is learned during the learning process [Jin *et al.*, 2025]. These methods require an absolute loss-value threshold. Otherwise, they fail to function effectively during training. Inspired by the principle of robust learning, ProAd-FS employs a ranking method to adjust loss values while maintaining their fundamental stability, providing a robust gate controller for feature discovery. These methods require a predefined absolute loss threshold, which is fragile in the volatile training dynamics of adversarial settings. Inspired by the principle of robust learning, ProAd-FS employs a ranking method to adjust loss values. This design is inherently invariant to monotonic shifts in the loss distribution, providing a stable gating mechanism for feature discovery.

3 Proposed Methodology

3.1 Overall Objective Function

Let a multi-view dataset be denoted by $\mathbf{X} = \{\mathbf{X}^{(v)} \in \mathbb{R}^{n\times d_v}\}_{v=1}^{V}$, where n is the number of samples and d_v is the feature dimensionality of the v-th view. Our goal is to learn a set of linear projections $\mathbf{W} = \{\mathbf{W}^{(v)} \in \mathbb{R}^{d_v\times k}\}_{v=1}^{V}$ such that the resulting k-dimensional latent representations $\mathbf{H}^{(v)} = \mathbf{X}^{(v)}\mathbf{W}^{(v)}$ are learned by a set of competing objectives: they should be sufficiently cross-view consistent to confuse a discriminator trained to detect their origin, while also being sufficiently intra-view diverse to appear indistinguishable from a prior distribution to another set of discriminators.

The ProAd-FS framework is formulated as a min-max optimization problem. The minimizer controls the projection matrices $\mathbf{W}$ and a sample-weighting vector $\mathbf{s}$, while the maximizer consists of two distinct discriminators: a consistency discriminator $\mathbf{D_c}$ tasked with identifying the origin view of latent samples, and a diversity discriminator $\mathbf{D_d}$ tasked with distinguishing these representations from synthetic prior distributions. To ensure the correctness of the learning trajectory, we introduce Robust Gated Curriculum (Section 3.3), which progressively selects easier-to-train samples. This process is implemented by the indicator function $\mathcal{I}_{\mathcal{C}_t}(\mathbf{s})$, with hard constraints imposed on the vector $\mathbf{s}$. The constraint set

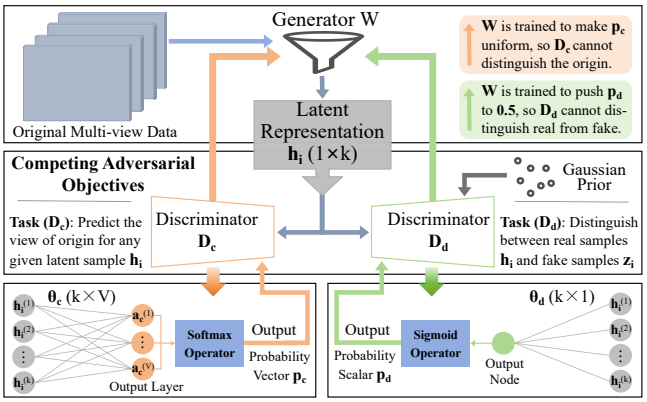


Figure 1: The adversarial architecture in ProAd-FS.

$\mathcal{C}_t = \{\mathbf{s} \in \{0,1\}^n \mid \|\mathbf{s}\|_1 = N_t\}$ specifies that exactly n_t samples must be selected during the t-th iteration. The sample budget N_t is not fixed but varies over time, enabling the model to learn from more complex samples. The overall objective function is defined as follows:

$$\min_{\mathbf{W},\mathbf{s}} \max_{\mathbf{D_c},\mathbf{D_d}} \Big(\sum_{i=1}^{n} s_i \mathcal{L}_i(\mathbf{W}, \mathbf{D_c}, \mathbf{D_d}) + \mathcal{I}_{\mathcal{C}_t}(\mathbf{s}) + \beta \sum_{v=1}^{V} \|\mathbf{W}^{(v)}\|_{2,1} \Big) \tag{1}$$

where $\mathcal{L}_i$ is the sample-level adversarial loss, while β is the hyperparameter that balances the loss with $\ell_{2,1}$-norm sparse regularization. We now define each component in detail.

3.2 Competing Adversarial Objectives

For each sample loss $\mathcal{L}_i$ that comprises two adversarial games, we didn't use a single objective function to balance their consistency and diversity. Instead, we assign these objectives to specific competition mechanisms to solve the problem. The relationship between the generator and discriminator is shown in Figure 1.

Inter-View Consistency Game. The task of the consistency game is to enforce distributional alignment between latent representations, leading to a view-invariant learned latent space. To achieve this, we design a multi-class consistency discriminator $\mathbf{D_c}$, parameterized by θ_c. This attempts to predict the view of origin for any given latent sample $\mathbf{h}$. The projection matrices $\mathbf{W}$ act as a single multi-part generator, trained to produce latent representations that are so similar across views that $\mathbf{D_c}$ cannot distinguish them. This is formulated as a Min-Max game over the negative log-likelihood of the correct view prediction:

$$\mathcal{L}_{\text{cons}}(\mathbf{W}, \mathbf{D_c}) = \min_{\mathbf{W}} \max_{\mathbf{D_c}} \mathbb{E}_{v\sim\mathcal{U}(1,V)} \Big[\mathbb{E}_{\mathbf{h}\sim P_{\mathbf{H}^{(v)}}} [\log \mathbf{D_c}(y = v|\mathbf{h})] \Big] \tag{2}$$

where $\mathbf{D_c}(y = v|\mathbf{h})$ denotes the discriminator's probability $\mathbf{p}_c$ that latent sample $\mathbf{h}$ belongs to view v. This is computed by the Softmax function from the logit score vector $\mathbf{a}_c = \mathbf{h}\theta_c$. The discriminator $\mathbf{D_c}$ aims to maximize this objective (correct classification), whereas the generator $\mathbf{W}$ aims to minimize it (induce misclassification).

Intra-View Diversity Game. In the diversity game within the view, to prevent all samples from being mapped to a trivial solution, the model ensures the latent representation to preserve the rich structure of the original data. For each view v, a binary diversity discriminator $\mathbf{D_d}^{(v)}$ is defined by a parameter $\theta_d^{(v)}$, which discriminates between "real" samples drawn from the latent distribution $\mathbf{h}^{(v)} \sim P_{\mathbf{H}^{(v)}}$ and "fake" samples drawn from the standard Gaussian prior distribution $p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I})$. The Min-Max game is governed by the cross-entropy loss function.

$$\mathcal{L}_{\text{div}}^{(v)}(\mathbf{W}^{(v)}, \mathbf{D_d}^{(v)}) = \min_{\mathbf{W}^{(v)}} \max_{\mathbf{D_d}^{(v)}} \Big(\mathbb{E}_{\mathbf{h}^{(v)}\sim P_{\mathbf{H}^{(v)}}} [\log \mathbf{D_d}^{(v)}(\mathbf{h}^{(v)})] + \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})} [\log(1 - \mathbf{D_d}^{(v)}(\mathbf{z}))] \Big) \tag{3}$$

The discriminator $\mathbf{D_d}$ is designed to estimate the probability scalar $\mathbf{p}_d$ that real samples are labeled as 1 while fake samples are labeled as 0. This scalar is generated by the Sigmoid function for the Logit score $\mathbf{a_d} = \mathbf{h}\theta_d^{(v)}$. The generator $\mathbf{W}^{(v)}$ minimizes this expression by transforming $\mathbf{X}^{(v)}$ to $\mathbf{H}^{(v)}$, with its distribution indistinguishable from $p(\mathbf{z})$.

Total Adversarial Loss. The total adversarial loss is computed with a single loss value $\mathcal{L}_i$ for each sample during the training process. From the generator's perspective, the algorithm achieves loss minimization through a weighted sum of consistency loss and diversity loss, as detailed in Algorithm 1. Instead of balancing fixed objectives, γ acts as a scaling factor for the gradient signals that arise from two adversarial games. This trade-off is dynamically adjusted by the relative strength of $\mathbf{D_c}$ and $\mathbf{D_d}$ throughout the training process.

Algorithm 1 Per-Sample Adversarial Loss Computation

Input: Sample $\mathbf{X}_i^{(v)}$; Projection matrix $\mathbf{W}^{(v)}$; Discriminators $\mathbf{D}_c, \mathbf{D}_d$; Scaling factor γ, Small epsilon ϵ.
Output: Per-sample loss $\mathcal{L}_i$.
$\mathbf{h}_i \leftarrow \mathbf{X}_i^{(v)} \mathbf{W}^{(v)}$
$\mathcal{L}_{\text{cons}}(h_i) \leftarrow -\log(1 - \mathbf{D}_c(y = v|\mathbf{h}_i) + \epsilon)$
$\mathcal{L}_{\text{div}}^{(v)}(h_i) \leftarrow -\log(1 - \mathbf{D}_d^{(v)}(\mathbf{h}_i) + \epsilon)$
$\mathcal{L}_i \leftarrow \mathcal{L}_{\text{cons}}(\mathbf{h}_i) + \gamma \mathcal{L}_{\text{div}}^{(v)}(\mathbf{h}_i)$
return $\mathcal{L}_i$

3.3 Robust Gated Curriculum

The RGC is designed to impose a hierarchical learning path to resist the volatile optimization dynamics in adversarial training. It consists of two interconnected components: gating to regulate training data volume and a robust selection criterion. In adversarial scenarios, features form a non-stationary loss function landscape where absolute loss values fluctuate during iterations. Therefore, content selection should prioritize samples with relative lower complexity over those exhibiting extreme volatility.

This strategy exhibits inherent robustness, as the monotonicity of the loss distribution remains unchanged. The gating mechanism consistently selects the simplest samples at

each stage, while progressively increasing the active sample size N_t with each iteration t, thereby achieving a deterministic annealing process. The transition from simple to complex samples is smooth, with the sample selection ratio δ_t determined by the following sigmoidal annealing function:

$$\delta_t = \delta_{\text{start}} + \frac{1 - \delta_{\text{start}}}{1 + \exp\left(-\sigma \cdot \frac{t - T/2}{T}\right)} \tag{4}$$

where t denotes the current iteration count, δ_{start} represents the initial sample proportion, T indicates the total iteration count, σ is the slope of the sigmoidal curve, and $N_t = \lfloor \delta_t \cdot n \rfloor$ is the sample size. For the step of calculating and selecting the easiest samples, see Algorithm 2.

Algorithm 2 Robust Gated Curriculum Update

Input: Samples $\{x_i\}_{i=1}^n$; Current iteration t; Total iterations T; Empty list $\mathcal{L}$.
Output: Curriculum weights $\mathbf{s}$.
$\delta_t \leftarrow$ Compute annealing percentage by Eq. (4);
$N_t \leftarrow \lfloor \delta_t \cdot n \rfloor$;
for $i = 1$ to n **do**
  $\mathcal{L}_i \leftarrow$ Compute per-sample loss using Algorithm 1;
  Append $(\mathcal{L}_i, i)$ to $\mathcal{L}$;
end for
Sort $\mathcal{L}$ based on loss values in ascending order;
Initialize sample weight vector $\mathbf{s} \leftarrow \mathbf{0} \in \mathbb{R}^n$;
for $k = 1$ to N_t **do**
  $(\sim, \text{idx}_k) \leftarrow \mathcal{L}[k]$;
  $\mathbf{s}_{\text{idx}_k} \leftarrow 1$;
end for
return $\mathbf{s}$

3.4 $\ell_{2,1}$-Norm Regularization

To achieve feature selection, we impose a group sparsity constraint on the rows of the projection matrices $\mathbf{W}^{(v)}$. The $\ell_{2,1}$-norm of a matrix $\mathbf{W}^{(v)}$ is defined as $\|\mathbf{W}^{(v)}\|_{2,1} = \sum_{j=1}^{d_v} \|\mathbf{w}_j^{(v)}\|_2$, where $\mathbf{w}_j^{(v)}$ is the j-th row vector [Nie *et al.*, 2010]. Minimizing this norm encourages entire rows of $\mathbf{W}^{(v)}$ to become zero, effectively reducing redundant features.

4 Optimization

The Min-Max objective function of ProAd-FS (Eq. 1) involves multiple interdependent variables and does not possess a closed-form solution. Therefore, we devise an iterative alternating optimization strategy to find a stable equilibrium.

4.1 Minimizer Optimization

Updating Sample Weights $\mathbf{s}$. Given the fixed generator $\mathbf{W}$ and discriminator $\mathbf{D}$, the constraint on the sample weights $\mathbf{s}$ from our overall objective function $\mathcal{J}$ (Eq. (1)) is a constrained optimization problem:

$$\min_{\mathbf{s}\in\{0,1\}^n, \|\mathbf{s}\|_1 = N_t} \sum_{i=1}^{n} s_i \mathcal{L}_i \tag{5}$$

where the sample loss function $\mathcal{L}_i$ measures sample complexity, and the optimal solution is provided by the following theorem.

Theorem 4.1. *Let $\{\mathcal{L}_j, idx_j\}_{j=1}^n$ be the sequence of per-sample losses and their original indices, sorted such that $\mathcal{L}_1 \le \mathcal{L}_2 \le \cdots \le \mathcal{L}_n$. The optimal solution $\mathbf{s}^*$ to the problem in Eq.* (5) *is given by:*

$$\mathbf{s}_i^* = \begin{cases} 1, & \text{if} \quad i \in \{idx_1, \ldots, idx_{N_t}\} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Proof. The proof is by a standard exchange argument. Assume there exists an optimal solution $\mathbf{s}'$ that is different from $\mathbf{s}^*$. This implies there must be at least one index $k \in \{\text{idx}_1, \ldots, \text{idx}_{N_t}\}$ for which $\mathbf{s}'_k = 0$, and at least one index $m \in \{\text{idx}_{N_t+1}, \ldots, \text{idx}_n\}$ for which $\mathbf{s}'_m = 1$. By the definition of our sorted losses, we have $\mathcal{L}_k \le \mathcal{L}_m$.

Let us construct a new solution $\mathbf{s}''$ by swapping these two assignments: let $\mathbf{s}''_k = 1$, $\mathbf{s}''_m = 0$, and $\mathbf{s}''_i = \mathbf{s}'_i$ for all other i. The new solution $\mathbf{s}''$ remains feasible, as $\|\mathbf{s}''\|_1 = \|\mathbf{s}'\|_1 = N_t$. The objective value for $\mathbf{s}''$ is $\sum_{i=1}^n \mathbf{s}''_i \mathcal{L}_i = (\sum_{i=1}^n \mathbf{s}'_i \mathcal{L}_i) - \mathcal{L}_m + \mathcal{L}_k$. Since $\mathcal{L}_k \le \mathcal{L}_m$, we have $\sum_{i=1}^n \mathbf{s}''_i \mathcal{L}_i \le \sum_{i=1}^n \mathbf{s}'_i \mathcal{L}_i$. Therefore, the objective value of $\mathbf{s}''$ is no worse than $\mathbf{s}'$. By repeating this exchange for all such differing pairs, we can transform $\mathbf{s}'$ into $\mathbf{s}^*$ without increasing the objective value. □

Updating Projection Matrices $\mathbf{W}$. With the curriculum $\mathbf{s}^*$ and the discriminators $\mathbf{D}$ fixed, we update the projection matrices $\mathbf{W}$ by performing a single step of gradient descent. The optimization of each projection matrix $\mathbf{W}^{(v)}$ is achieved by descending along the gradient of $\mathcal{J}$:

$$\nabla_{\mathbf{W}^{(v)}} \mathcal{J} = \nabla_{\mathbf{W}^{(v)}} \mathcal{J}_{\text{cons}} + \gamma \nabla_{\mathbf{W}^{(v)}} \mathcal{J}_{\text{div}} + \beta \nabla_{\mathbf{W}^{(v)}} \mathcal{J}_{\text{reg}} \tag{7}$$

where each component of the gradient is a weighted sum over the samples, detailed as follows:

$$\begin{cases} \nabla_{\mathbf{W}^{(v)}} \mathcal{J}_{\text{cons}} = (\mathbf{X}^{(v)})^T (\mathbf{s}^* \odot \nabla_{\mathbf{H}^{(v)}} \mathcal{L}_{\text{cons}}) \\ \nabla_{\mathbf{W}^{(v)}} \mathcal{J}_{\text{div}} = (\mathbf{X}^{(v)})^T (\mathbf{s}^* \odot \nabla_{\mathbf{H}^{(v)}} \mathcal{L}_{\text{div}}^{(v)}) \\ \nabla_{\mathbf{W}^{(v)}} \mathcal{J}_{\text{reg}} = \Lambda^{(v)} \mathbf{W}^{(v)} \end{cases} \tag{8}$$

where $\odot$ is the Hadamard product. The diagonal matrix $\Lambda^{(v)}$ contains the inverse ℓ_2-norms of the rows of $\mathbf{W}^{(v)}$.

Aggregating these components, the update rule for the projection matrices is executed with a learning rate η_w:

$$\mathbf{W}^{(v)} \leftarrow \mathbf{W}^{(v)} - \eta_w \nabla_{\mathbf{W}^{(v)}} \mathcal{J} \tag{9}$$

4.2 Maximizer Optimization

Updating Consistency Discriminator $\mathbf{D_c}$. The parameters $\boldsymbol{\theta}_c$ are updated using the gradient of the weighted cross-entropy loss. The gradient is derived from the prediction error $(\mathbf{Y}_c - \mathbf{P}_c)$, where $\mathbf{Y}_c$ represents the one-hot view indices and $\mathbf{P}_c$ the predicted probabilities. This error is scaled by the curriculum weights $\mathbf{s}^*$, and the resulting weighted error is backpropagated. The update rule is:

$$\boldsymbol{\theta}_c \leftarrow \boldsymbol{\theta}_c + \eta_D \nabla_{\boldsymbol{\theta}_c} \mathcal{J}_{\text{cons}} \tag{10}$$

where $\nabla_{\boldsymbol{\theta}_c} \mathcal{J}_{\text{cons}}$ is proportional to $(\mathbf{H}_c)^T (\mathbf{s}^* \odot (\mathbf{Y}_c - \mathbf{P}_c))$.

Updating Diversity Discriminators $\mathbf{D_d}^{(v)}$. The update for the view-specific parameters $\boldsymbol{\theta}_d^{(v)}$ relies on a composite batch $\mathbf{H}_d$ and an asymmetric weighting strategy. The batch $\mathbf{H}_d$ comprises both real samples in the form of latent representations $\mathbf{H}^{(v)}$ and fake samples consisting of synthetic vectors $\mathbf{z}$ from a Gaussian prior. The corresponding authenticity labels $\mathbf{Y}_d$ are generated on-the-fly (1 for real, 0 for fake). Crucially, the curriculum is applied asymmetrically, where the prediction error for real samples is weighted by $\mathbf{s}^*$ in contrast to the error for fake samples, which is uniformly weighted by one. This is mathematically realized by constructing a composite weight vector $\mathbf{s}_d^* = [\mathbf{s}^*; \mathbf{1}]$. This design fosters a more robust adversary by compelling it to learn the entire prior distribution. The update rule is:

$$\boldsymbol{\theta}_d^{(v)} \leftarrow \boldsymbol{\theta}_d^{(v)} + \eta_D \nabla_{\boldsymbol{\theta}_d^{(v)}} \mathcal{J}_{\text{div}} \tag{11}$$

where $\nabla_{\boldsymbol{\theta}_d^{(v)}} \mathcal{J}_{\text{div}}$ is proportional to $(\mathbf{H}_d)^T(\mathbf{s}_d^* \odot (\mathbf{Y}_d - \mathbf{P}_d))$, reflecting the backpropagation of weighted error.

5 Computational Complexity

We analyze the computational complexity of the ProAd-FS algorithm per iteration. The forward pass commences with the projection of input data into the latent space. For each of the V views, this involves a matrix multiplication over all views between $\mathbf{X}^{(v)}$ and $\mathbf{W}^{(v)}$, incurring a cost of $\mathcal{O}(\sum_{v=1}^{V} nd_v k)$. Subsequently, computing per-sample losses is dominated by the consistency discriminator, which has a complexity of $\mathcal{O}(V^2 nk)$. The curriculum update then requires a sorting operation on the n loss values, which costs $\mathcal{O}(n \log n)$. During the backward pass, the gradients for the projection matrices are computed. For each view v, it requires a multiplication of the form $(\mathbf{X}^{(v)})^\top(\cdot)$, where a matrix of size $d_v \times n$ is multiplied by a matrix of size $n \times k$. The cost for all V views is therefore $\mathcal{O}(\sum_{v=1}^{V} d_v nk)$.

Aggregating these costs, the total complexity per iteration is $\mathcal{O}\left(\sum_{v=1}^{V} d_v nk + V^2 nk + n \log n\right)$. In typical large-scale scenarios where multi-view data is used, the feature dimensions d_v are significantly larger than the number of views V and $\log n$. The term $\sum_{v=1}^{V} d_v nk$ dominates the others. As a result, the per-sample total is approximately $\mathcal{O}\left(nk \sum_{v=1}^{V} d_v\right)$. Therefore, the complexity of ProAd-FS is linear with respect to both the number of samples and the total feature dimensionality.

6 Experiments

6.1 Experimental Setup

Datasets. We selected six multi-view datasets that span diverse types and scales to ensure a comprehensive evaluation, including MSRC-v1[1], ORL[2], COIL20[3], Animal[4], ALOI[5], and FashionMNIST[6]. Detailed statistics for each dataset are summarized in Table 1.

Dataset	Samples	Classes	Features
MSRC-v1	210	7	{1302, 48, 512, 256, 210}
ORL	400	40	{4096, 3304, 6750}
COIL20	1440	20	{944, 324, 512}
Animal	10158	50	{4096, 4096}
ALOI	10800	100	{77, 13, 64, 125}
FashionMNIST	70000	10	{784, 256, 6084, 2304, 59, 500}

Table 1: Details of datasets.

Comparison Methods. We compared ProAd-FS against a variety of baselines covering a wide spectrum of UMFS methods. Our comparisons included advanced multi-view methods such as JMVFG [Fang *et al.*, 2023a], CAHR [Song *et al.*, 2024], CE-UMFS [Zhou and Song, 2024], and UKMFS [Hu *et al.*, 2025]. We also included a leading single-view method adapted for this task (FSDK [Nie *et al.*, 2023]) and a full-feature baseline (AllFea) to establish an initial performance floor. The theoretical complexity of the methods is presented in Table 2.

Method	Main Time Complexity (per iteration)
FSDK	$\mathcal{O}\left(d^3 + nd^2\right)$
JMVFG	$\mathcal{O}\left(\sum_{v=1}^{V}(d_v^3 + nd_v^2 + n^2 d_v)\right)$
CAHR	$\mathcal{O}\left(n^2 \log(n) + nd^3\right)$
CE-UMFS	$\mathcal{O}\left(n\sum_{v=1}^{V} d_v^2 + Vn^2 k\right)$
UKMFS	$\mathcal{O}\left(Vn^3 + Vnd^2 + Vd^3\right)$
ProAd-FS	$\mathcal{O}\left(nk\sum_{v=1}^{V} d_v\right)$

Table 2: Comparison of per-iteration computational complexity.

Implementation Details. Following standard practice, we applied each method to select a feature subset. We then evaluated the quality of the selected features by performing K-means clustering and reported three standard metrics: Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI) [Xu and Wunsch, 2005]. To ensure statistical significance, we averaged all results over 20 independent runs, and conducted statistical comparisons in accordance with the recommendations of [Zhou *et al.*, 2025]. Furthermore, we tuned γ and β via a grid search over $\{0.01, 0.1, 1, 10, 100\}$, and set the learning rates to $\eta_w = 10^{-4}$ and $\eta_D = 10^{-5}$. Finally, we scheduled the curriculum to complete at 40% of the total iterations.

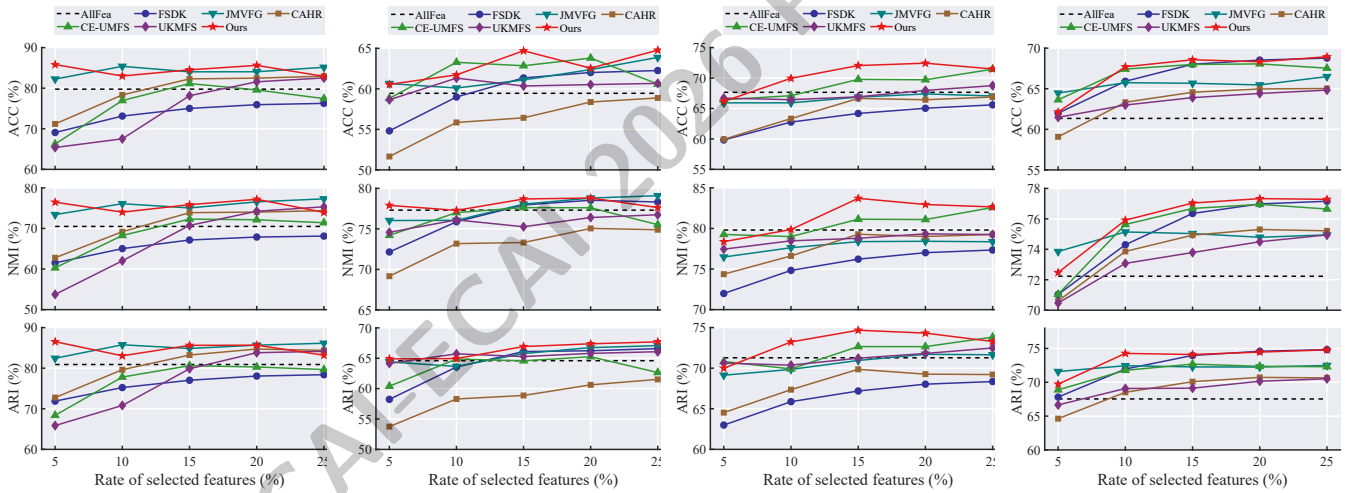
6.2 Clustering Results

To validate the effectiveness of ProAd-FS, we conducted a clustering performance evaluation. Figure 2 demonstrates that our ProAd-FS framework consistently outperforms all baseline methods across four representative datasets, even surpassing the full dataset at low selection rates (5-10%). This highlights the effectiveness of our adversarial game in identifying discriminative feature subsets.

[1] http://www.cad.zju.edu.cn/home/dengcai/Data/data.html
[2] https://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html
[3] https://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php
[4] https://opendatalab.org.cn/OpenDataLab/ANIMAL
[5] https://aloi.science.uva.nl/
[6] https://www.worldlink.com.cn/en/osdir/fashion-mnist.html

Dataset	Metric	Method						
		AllFea	FSDK	JMVFG	CAHR	CE-UMFS	UKMFS	ProAd-FS
MSRC-v1	ACC	79.76±5.51	75.01±3.07	84.03±6.67	82.28±5.32	81.14±3.13	78.14±4.02	**84.52±4.93**†
	NMI	70.49±5.27	67.15±3.85	75.05±5.59	73.88±4.72	72.30±2.84	70.74±3.04	**75.86±4.61**
	ARI	80.91±4.46	77.02±2.31	84.85±4.98	83.23±3.54	80.61±2.79	79.85±3.58	**85.62±5.04**†
ORL	ACC	59.45±4.30	61.35±2.53	61.10±2.86	56.42±2.84	62.85±3.07	60.35±2.28	**64.70±3.89**†
	NMI	77.31±2.19	77.96±1.57	78.05±1.12	73.31±1.31	77.62±1.49	75.26±1.06	**78.69±2.35**†
	ARI	64.60±3.42	66.10±2.37	65.78±1.91	58.87±1.63	64.57±2.57	65.27±2.01	**66.93±2.74**†
COIL20	ACC	67.65±3.35	64.18±3.60	66.86±3.51	66.65±2.89	69.77±2.47	66.91±2.14	**72.05±3.82**†
	NMI	79.81±2.26	76.21±1.97	78.37±1.48	79.28±1.03	81.13±1.72	78.79±0.79	**83.71±2.77**†
	ARI	71.28±3.07	67.18±3.42	70.95±1.78	69.86±1.90	72.66±1.64	71.23±1.54	**74.67±2.83**†
Animal	ACC	61.34±1.22	68.02±1.92	65.67±1.61	64.56±2.02	67.99±2.24	63.91±2.16	**68.57±2.09**†
	NMI	72.23±0.61	76.35±0.31	75.03±0.45	74.93±0.59	76.68±0.69	73.78±0.67	**77.04±0.65**†
	ARI	67.53±1.29	73.96±1.13	72.27±0.67	70.08±1.08	72.69±1.86	69.12±1.49	**74.10±1.48**†
ALOI	ACC	73.16±2.70	75.60±1.77	**77.54±2.83**	76.81±1.65	76.49±1.33	76.60±2.01	77.17±2.68
	NMI	92.28±0.82	92.94±0.59	93.26±0.66	93.12±0.48	93.01±0.34	93.29±0.54	**93.59±0.69**†
	ARI	78.67±1.98	83.29±1.69	83.07±2.13	83.20±1.41	82.54±0.90	81.70±1.73	**83.68±2.03**†
FashionMNIST	ACC	54.57±0.01	53.74±0.37	N/A	N/A	56.71±0.64	N/A	**58.64±0.03**†
	NMI	59.66±0.01	55.53±0.71	N/A	N/A	58.49±0.34	N/A	**61.91±0.01**†
	ARI	59.71±0.01	56.82±0.83	N/A	N/A	61.38±0.67	N/A	**62.41±0.03**†

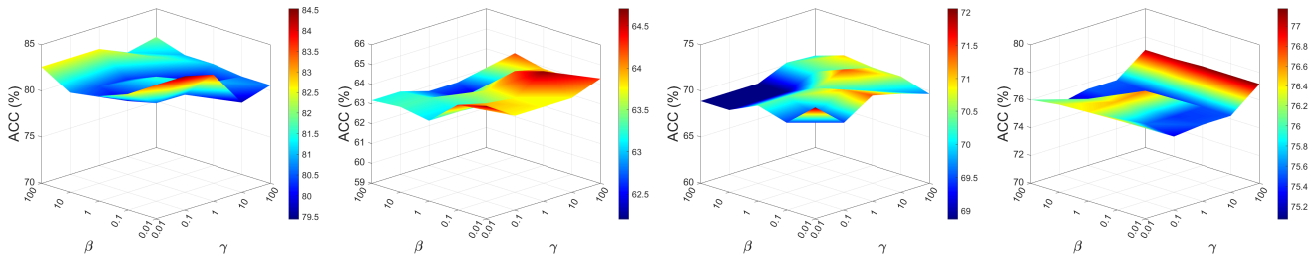
Table 3: Comparative clustering performance (mean±std%) using K-means clustering across all datasets. Entries marked as "N/A" could not be computed due to excessive memory requirements. Best results are highlighted in **bold**, and second-best results are underlined. The symbol † indicates statistical significance ($p < 0.05$) compared to the runner-up using a paired t-test.



(a) MSRC-v1 (b) ORL (c) COIL20 (d) Animal

Figure 2: Clustering performance of different feature selection algorithms across varying selection rates.

Table 3 details the performance at a more reasonable 15% feature selection rate. Our method still demonstrates significant advantages in the majority of cases. The statistical significance test results (marked with †) indicate a clear advantage of our approach over suboptimal methods ($p < 0.05$). Specifically, on the COIL20 dataset, our method achieves a distinct advantage, showing the efficacy of the decoupled adversarial architecture. On the ALOI dataset, while our method does not achieve the highest accuracy, it excels in both NMI and ARI metrics. This is because NMI and ARI are key indicators for evaluating clustering structure fidelity. Our framework preserves the inherent data structure, which proves more valuable than simple label matching accuracy in certain unsupervised scenarios. The scalability advantage of ProAd-FS was validated on the large-scale FashionMNIST dataset. Compared to the multi-view baseline methods rendered "N/A" (Not Available) by their super-linear complexity, ProAd-FS maintains computational feasibility while achieving the highest scores across all metrics. These empirical results validate both the effectiveness and the scalable superiority of ProAd-FS.



(a) MSRC-v1 (b) ORL (c) COIL20 (d) ALOI

Figure 3: Parameter sensitivity analysis w.r.t. γ and β on four datasets. The z-axis and color map represent the values of ACC.

6.3 Running Time

To validate the efficiency of the ProAd-FS framework, we measured the runtime of various baselines across four representative datasets. Table 4 demonstrates its practical benefits in linear complexity design, where ProAd-FS exhibits superior performance, especially on challenging datasets. On the ORL dataset, ProAd-FS is significantly faster than methods like CAHR, which have high complexity for feature dimensions. On the large-scale FashionMNIST dataset, while other competing methods fail due to Out-of-Memory (OM) errors, ProAd-FS operates with significantly higher efficiency. The experimental results aligned closely with our theoretical complexity analysis outlined in Table 2. The linear-complexity architecture of ProAd-FS is not just a theoretical benefit, but translates directly to practical implementations, confirming its efficiency and scalability for high-dimensional and large-scale applications.

Method	ORL	Animal	ALOI	FashionMNIST
FSDK	234.8	166.4	120.8	5293.6
JMVFG	55.0	695.4	773.1	N/A
CAHR	313.4	2997.3	3072.4	N/A
CE-UMFS	76.9	569.4	308.5	9633.4
UKMFS	58.1	437.3	404.1	N/A
ProAd-FS	**33.2**	**103.2**	**90.3**	**290.1**

Table 4: Comparison of runtime (in seconds).

6.4 Ablation Study

To validate the necessity of our comprehensive design, we employed three variants of ProAd-FS: 1) without Consistency (w/o Cons.), which removes the inter-view consistency game; 2) without Diversity (w/o Div.), which removes the intra-view diversity game; and 3) without Robust Gated Curriculum (w/o RGC), which removes the Robust Gated Curriculum. As shown in Table 5, removing both consistency and diversity components resulted in performance degradation across all datasets. This occurs because the model cannot learn core shared information across views when alignment targets are absent, which is fundamental to multi-view learning. The significant performance drop caused by removing the RGC also demonstrates its critical role in the optimization process. This is because the RGC establishes the learning process in a progressive complexity order, which is crucial for achieving a high-quality solution. These three components collectively validate the effectiveness of our complete design.

Dataset	Metric	w/o Cons.	w/o Div.	w/o RGC	Full
MSRC-v1	ACC	77.4 (↓ 7.1)	80.4 (↓ 4.1)	78.8 (↓ 5.7)	84.5
	NMI	69.7 (↓ 6.2)	72.2 (↓ 3.7)	70.8 (↓ 5.1)	75.9
	ARI	78.7 (↓ 6.9)	81.2 (↓ 4.4)	80.2 (↓ 5.4)	85.6
COIL20	ACC	66.8 (↓ 5.3)	69.5 (↓ 2.6)	68.0 (↓ 4.1)	72.1
	NMI	78.0 (↓ 5.7)	80.9 (↓ 2.8)	79.9 (↓ 3.8)	83.7
	ARI	70.2 (↓ 4.5)	71.7 (↓ 3.0)	70.5 (↓ 4.2)	74.7
ALOI	ACC	72.3 (↓ 4.9)	75.6 (↓ 1.6)	73.6 (↓ 3.6)	77.2
	NMI	91.7 (↓ 1.9)	92.7 (↓ 0.9)	91.9 (↓ 1.7)	93.6
	ARI	78.5 (↓ 5.2)	81.7 (↓ 2.0)	79.5 (↓ 4.2)	83.7

Table 5: Ablation study on three datasets. Values in (↓ ·) indicate the performance degradation compared to the full ProAd-FS framework.

6.5 Hyperparameter Robustness

We investigated the sensitivity of our ProAd-FS framework to its two main hyperparameters: the scaling factor γ and the sparsity regularization weight β. Figure 3 shows the ACC for varying pairs of these two parameters for the four benchmark datasets. As can be seen, the results reveal a large contiguous plateau of high performance across all datasets, rather than an isolated peak. This indicates that ProAd-FS is not sensitive to the exact choice of hyperparameters and does not require detrimental exhaustive fine-tuning. In fact, significant degradation in performance only occurs at the extremes of the grid investigated for the parameter values where the objectives are highly imbalanced. For our experiments, optimal parameters were easily found via a standard grid search over the set $\{0.01, 0.1, 1, 10, 100\}$.

7 Conclusions

This work introduced a novel ProAd-FS framework to overcome the limitations in modeling the competition between cross-view alignment and intra-view diversity. We formulated this static trade-off as a dynamic competitive process in an adversarial decoupled architecture, where these competing objectives are assigned to distinct games. To ensure a principled learning trajectory within this adversarial setting, we introduced the RGC strategy that guides the model from underlying shared structures to complex view-specific information. An important aspect of ProAd-FS is that we achieve linear computational complexity in both sample size and feature dimension. This makes ProAd-FS suitable for large-scale applications, which would presumably be a limiting factor in existing UMFS methods. In summary, ProAd-FS unites a game-theoretic formulation with a highly efficient implementation, establishing an effective new solution to modern large-scale UMFS.

Acknowledgments

This work was supported in part by Guangdong Basic and Applied Basic Research Foundation (NO. 2024A1515030022), National Natural Science Foundation of China (No. 62276103, 61906069), Innovation Team Project of General Colleges and Universities in Guangdong Province (No. 2023KCXTD002), the Fundamental Research Funds for the Central Universities, JLU (No. 93K172024K03, 93K172024K24, 2024ZYGXZR097), the Natural Science Research Project of Education Department of Guizhou Province (No. QJJ2023061), Analysis of Urban Events Based on Low-Altitude Drones (No. 2024BQ010011), the Research and Development Project on Key Technologies for Intelligent Sensing, Guangdong Provincial Philosophy and Social Sciences Planning 2023 Key Project on Special Research in Auditing Theory (No. GD23SJZ09).

References

[Bengio *et al.*, 2009] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 41–48. PMLR, 2009.

[Cao and Xie, 2024] Zhiwen Cao and Xijiong Xie. Partition-level tensor learning-based multiview unsupervised feature selection. *IEEE Transactions on Neural Networks and Learning Systems*, 36(7):12799–12811, 2024.

[Chen *et al.*, 2016] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29. Curran Associates, Inc., 2016.

[Donahue and Simonyan, 2019] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc., 2019.

[Dong *et al.*, 2025] Xia Dong, Danyang Wu, Yiping Ke, and Feiping Nie. Unsupervised multi-view feature selection via attentive hierarchical bipartite graphs with optimizable graph filter. *Information Fusion*, 126:103511, 2025.

[Fadadu *et al.*, 2022] Sudeep Fadadu, Shreyash Pandey, Darshan Hegde, Yi Shi, Fang-Chieh Chou, Nemanja Djuric, and Carlos Vallespi-Gonzalez. Multi-view fusion of sensor data for improved perception and prediction in autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2349–2357, 2022.

[Fang *et al.*, 2023a] Si-Guo Fang, Dong Huang, Chang-Dong Wang, and Yong Tang. Joint multi-view unsupervised feature selection and graph learning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(1):16–31, 2023.

[Fang *et al.*, 2023b] Uno Fang, Man Li, Jianxin Li, Longxiang Gao, Tao Jia, and Yanchun Zhang. A comprehensive survey on multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12350–12368, 2023.

[Feng *et al.*, 2024] Shikun Feng, Yuyan Ni, Minghao Li, Yanwen Huang, Zhi-Ming Ma, Wei-Ying Ma, and Yanyan Lan. Unicorn: A unified contrastive learning approach for multi-view molecular representation learning. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235, pages 13256–13277. PMLR, 2024.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27. Curran Associates, Inc., 2014.

[Hou *et al.*, 2017] Chenping Hou, Feiping Nie, Hong Tao, and Dongyun Yi. Multi-view unsupervised feature selection with adaptive similarity and view weight. *IEEE Transactions on Knowledge and Data Engineering*, 29(9):1998–2011, 2017.

[Hsieh *et al.*, 2019] Tsung-Yu Hsieh, Yiwei Sun, Suhang Wang, and Vasant Honavar. Adaptive structural co-regularization for unsupervised multi-view feature selection. In *2019 IEEE International Conference on Big Knowledge (ICBK)*, pages 87–96. IEEE, 2019.

[Hu *et al.*, 2025] Rongyao Hu, Jiangzhang Gan, Mengmeng Zhan, Li Li, and Mengling Wei. Unsupervised kernel-based multi-view feature selection with robust self-representation and binary hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17287–17294, 2025.

[Jiang *et al.*, 2014] Lu Jiang, Deyu Meng, Shoou-I Yu, Zhenzhong Lan, Shiguang Shan, and Alexander G Hauptmann. Self-paced learning with diversity. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27. Curran Associates, Inc., 2014.

[Jin *et al.*, 2025] Jianhui Jin, Qiuping Jiang, Qingyuan Wu, Binwei Xu, and Runmin Cong. Underwater salient object detection via dual-stage self-paced learning and depth emphasis. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(3):2147–2160, 2025.

[Khalafaoui *et al.*, 2025] Yasser Khalafaoui, Basarab Matei, Martino Lovisetto, and Nistor Grozavu. Deep matrix factorization with adaptive weights for multi-view clustering. *Pattern Recognition*, 170:112027, 2025.

[Kong *et al.*, 2021] Yajing Kong, Liu Liu, Jun Wang, and Dacheng Tao. Adaptive curriculum learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5067–5076, 2021.

[Kumar *et al.*, 2010] M Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 23. Curran Associates, Inc., 2010.

[Li *et al.*, 2018] Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 31(10):1863–1883, 2018.

[Li *et al.*, 2022] Weiyi Li, Hongmei Chen, Tianrui Li, Jihong Wan, and Binbin Sang. Unsupervised feature selection via self-paced learning and low-redundant regularization. *Knowledge-Based Systems*, 240:108150, 2022.

[Liu *et al.*, 2025] Qi Liu, Suyuan Liu, Xinwang Liu, and Jianhua Dai. Latent semantics and anchor graph multilayer learning for multi-view unsupervised feature selection. *IEEE Transactions on Knowledge and Data Engineering*, 37(10):6032–6045, 2025.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 23. Curran Associates, Inc., 2010.

[Nie *et al.*, 2023] Feiping Nie, Zhenyu Ma, Jingyu Wang, and Xuelong Li. Fast sparse discriminative k-means for unsupervised feature selection. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):9943–9957, 2023.

[Pan *et al.*, 2021] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11205–11214, 2021.

[Song *et al.*, 2024] Peng Song, Shixuan Zhou, Jinshuai Mu, Meng Duan, Yanwei Yu, and Wenming Zheng. Clean affinity matrix induced hyper-laplacian regularization for unsupervised multi-view feature selection. *Information Sciences*, 682:121276, 2024.

[Tang *et al.*, 2013] Jiliang Tang, Xia Hu, Huiji Gao, and Huan Liu. Unsupervised feature selection for multi-view data in social media. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 270–278. SIAM, 2013.

[Tang *et al.*, 2021] Chang Tang, Xiao Zheng, Xinwang Liu, Wei Zhang, Jing Zhang, Jian Xiong, and Lizhe Wang. Cross-view locality preserved diversity and consensus learning for multi-view unsupervised feature selection. *IEEE Transactions on Knowledge and Data Engineering*, 34(10):4705–4716, 2021.

[Wan *et al.*, 2020] Yuan Wan, Shengzi Sun, and Cheng Zeng. Adaptive similarity embedding for unsupervised multi-view feature selection. *IEEE Transactions on Knowledge and Data Engineering*, 33(10):3338–3350, 2020.

[Wang *et al.*, 2021] Qi Wang, Xu Jiang, Mulin Chen, and Xuelong Li. Autoweighted multiview feature selection with graph optimization. *IEEE Transactions on Cybernetics*, 52(12):12966–12977, 2021.

[Wu *et al.*, 2024] Jian-Sheng Wu, Yanlan Li, Jun-Xiao Gong, and Weidong Min. Collaborative and discriminative subspace learning for unsupervised multi-view feature selection. *Engineering Applications of Artificial Intelligence*, 133:108145, 2024.

[Xu and Wunsch, 2005] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.

[Xu *et al.*, 2019a] Jinglin Xu, Junwei Han, Feiping Nie, and Xuelong Li. Multi-view scaling support vector machines for classification and feature selection. *IEEE Transactions on Knowledge and Data Engineering*, 32(7):1419–1430, 2019.

[Xu *et al.*, 2019b] Tianyang Xu, Zhen-Hua Feng, Xiao-Jun Wu, and Josef Kittler. Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking. *IEEE Transactions on Image Processing*, 28(11):5596–5609, 2019.

[Xu *et al.*, 2024] Cai Xu, Jiajun Si, Ziyu Guan, Wei Zhao, Yue Wu, and Xiyue Gao. Reliable conflictive multi-view learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16129–16137, 2024.

[Zhang *et al.*, 2024a] Chenglong Zhang, Yang Fang, Xinyan Liang, Han Zhang, Peng Zhou, Xingyu Wu, Jie Yang, Bingbing Jiang, and Weiguo Sheng. Efficient multi-view unsupervised feature selection with adaptive structure learning and inference. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[Zhang *et al.*, 2024b] Chenglong Zhang, Xinyan Liang, Peng Zhou, Zhaolong Ling, Yingwei Zhang, Xingyu Wu, Weiguo Sheng, and Bingbing Jiang. Scalable multi-view unsupervised feature selection with structure learning and fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 5479–5488, 2024.

[Zhou and Song, 2024] Shixuan Zhou and Peng Song. Consistency–exclusivity guided unsupervised multi-view feature selection. *Neurocomputing*, 569:127119, 2024.

[Zhou *et al.*, 2023a] Shixuan Zhou, Peng Song, Zihao Song, and Liang Ji. Soft-label guided non-negative matrix factorization for unsupervised feature selection. *Expert Systems with Applications*, 216:119468, 2023.

[Zhou *et al.*, 2023b] Shixuan Zhou, Peng Song, Yanwei Yu, and Wenming Zheng. Structural regularization based discriminative multi-view unsupervised feature selection. *Knowledge-Based Systems*, 272:110601, 2023.

[Zhou *et al.*, 2025] Shixuan Zhou, Yi Xiang, Han Huang, Pei Huang, Chaoda Peng, Xiaowei Yang, and Peng Song. Unsupervised feature selection with evolutionary sparsity. *Neural Networks*, 189:107512, 2025.

[Zuo *et al.*, 2023] Qiankun Zuo, Ning Zhong, Yi Pan, Huisi Wu, Baiying Lei, and Shuqiang Wang. Brain structure-function fusing representation learning using adversarial decomposed-vae for analyzing mci. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:4017–4028, 2023.