Wasserstein-Aligned Hyperbolic Multi-View Clustering

Rui Wang[1], **Yuting Jiang**[1], **Xiaoqing Luo**[1], **Xiao-Jun Wu**[1], **Nicu Sebe**[2], **Ziheng Chen**[2*]

[1]School of Artificial Intelligence and Computer Science, Jiangnan University
[2]Department of Information Engineering and Computer Science, University of Trento
{cs_wr, xqluo, wu_xiaojun}@jiangnan.edu.cn, {yuting_jiang2025, ziheng_ch}@163.com

Abstract

Multi-view clustering (MVC) aims to uncover the latent structure of multi-view data by learning view-common and view-specific information. Although recent studies have explored hyperbolic representations for better tackling the representation gap between different views, they focus primarily on instance-level alignment and neglect global semantic consistency, rendering them vulnerable to view-specific information (*e.g.*, noise and cross-view discrepancies). To this end, this paper proposes a novel Wasserstein-Aligned Hyperbolic (WAH) framework for multi-view clustering. Specifically, our method exploits a view-specific hyperbolic encoder for each view to embed features into the Lorentz manifold for hierarchical semantic modeling. Whereafter, a global semantic loss based on the hyperbolic sliced-Wasserstein distance is introduced to align manifold distributions across views. This is followed by soft cluster assignments to encourage cross-view semantic consistency. Extensive experiments on multiple benchmarking datasets show that our method can achieve SOTA clustering performance.

Code — https://github.com/Yuting-jiang-jnu/WAH-MVC

Introduction

Multi-view data consists of heterogeneous features or originates from multiple sources. Although describing the same underlying semantics, each view provides complementary information, capturing different aspects of the data. Integrating multiple views can significantly boost clustering performance (Chen et al. 2024; Xu et al. 2022a; Guo, Zhao, and Wang 2024). To this end, multi-view clustering (MVC) aims to exploit both the consistency and complementarity among views to achieve more accurate clustering. Traditional MVC approaches fall into several families, including subspace learning (Chen et al. 2022; Tao et al. 2021; Xie et al. 2024), nonnegative matrix factorization (Hu and Chen 2019; Wei et al. 2020), graph-based methods (Li, Wan, and He 2023; Chen et al. 2024), and kernel fusion (Liu et al. 2019, 2021). While effective in low-dimensional scenarios, these shallow models struggle with scalability and complex data due to limited representational power.

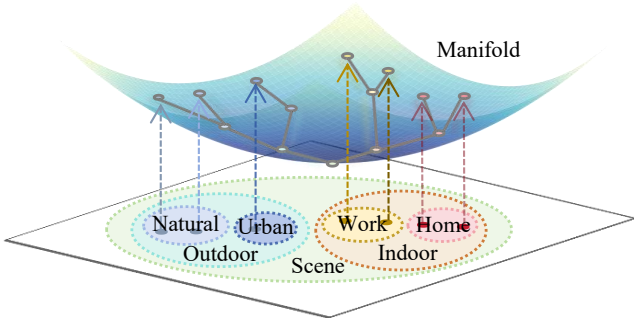


Figure 1: Real-world scene categories often form latent hierarchies. Here, fine-grained labels (*e.g.*, Home, Work, Natural, Urban) are grouped into higher-level concepts like Indoor and Outdoor. Hyperbolic space has shown success in modeling hierarchical structure.

To overcome these limitations, deep multi-view clustering (DMVC) methods have emerged as a promising paradigm that harnesses the representation power of deep neural networks to learn the latent embeddings from multiple views (Li et al. 2019, 2022; Wang et al. 2023; Zhang et al. 2024; Guo et al. 2024). These methods aim to simultaneously capture view-specific semantics while jointly discovering a shared clustering structure in the latent space. Xu et al. (2022a) introduce a self-supervised autoencoder with view consensus regularization, while Zhang et al. (2024) impose hierarchical latent constraints to enhance semantic consistency. However, reconstruction-based models may struggle to learn discriminative representations. To address this issue, Wang et al. (2023) employs adversarial strategies, combining view-specific encoders with discriminators to tackle this issue. However, adversarial training remains challenging due to its inherent instability and sensitivity, particularly in scenarios where the quality or completeness of views varies significantly (Xing, Song, and Cheng 2021; Xiao et al. 2022).

In contrast, contrastive learning has emerged as a more direct and stable alternative to enforce cross-view alignment (Peng et al. 2022; Li et al. 2022; Cui et al. 2024; Zhang et al. 2025; Hu et al. 2025). It encourages semantically similar samples from different views to reside nearby

*Corresponding author.

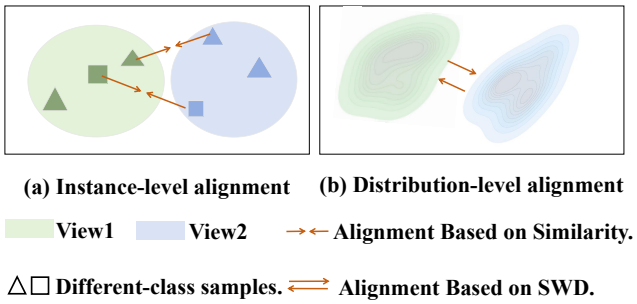


Figure 2: Comparison of traditional instance-level alignment and the proposed hyperbolic distribution-level alignment. Instance-level alignment maximizes pairwise similarity but poorly captures global semantics, whereas our method aligns holistic view-wise distributions via sliced-Wasserstein distance (SWD) on the Lorentz manifold.

in the latent space, while preserving the diversity of view-specific representations. Notably, Peng et al. (2022) disentangles view-invariant and view-specific components via contrastive objectives, whereas Cui et al. (2024) develops a dual-level contrastive framework that jointly aligns instance-level and cluster-level semantics. However, these methods are typically confined to Euclidean geometry, which limits their ability to capture non-Euclidean structures such as the hierarchical structure inherent in many real-world datasets (as illustrated in Figure 1).

In response, recent efforts have extended contrastive learning to non-Euclidean spaces to better capture the underlying geometric structures of complex data. Lin et al. (2022, 2023), for instance, embed features onto hyperbolic manifolds to model latent semantics with greater geometric fidelity. However, existing methods align views at the instance level, aiming to pull corresponding samples from different views closer while pushing apart unrelated ones based on feature similarity, as shown in Figure 2a. Although such a strategy is effective in aligning individual data points, it often overlooks the global semantic consistency across views. Moreover, it relies on manually selected positive and negative pairs, which introduces sampling bias and limits scalability when dealing with data containing a wide range of variations. To address these limitations, Optimal Transport (OT) has been introduced as a powerful tool for aligning global feature distributions under different views (Zhang et al. 2024). Unlike pointwise methods, distribution-level alignment via OT, most commonly quantified by the Wasserstein distance, captures holistic semantic correspondences by matching entire sample distributions. This approach effectively mitigates view gaps and enhances the learning of cross-view common semantics. However, generalizing Wasserstein distance to non-Euclidean geometries remains a challenging problem, as it requires preserving the intrinsic geometric structure of the data while maintaining computational efficiency.

In this paper, we propose WAH-MVC, a unified Wasserstein-Aligned Hyperbolic framework for multi-view clustering that simultaneously preserves hierarchical structures, aligns global manifold distributions, and enhances feature discriminability. Unlike prior methods based on the Poincaré model, we adopt the Lorentz model for its closed-form Riemannian operations and superior numerical robustness, which facilitate stable optimization and seamless integration with deep learning frameworks (Nickel and Kiela 2018; Chami et al. 2020). Specifically, WAH-MVC leverages multiple hyperbolic autoencoders to extract view-specific hierarchies and introduces a hyperbolic sliced-Wasserstein distance (SWD)-based constraint to achieve global distribution alignment across views in a geometry-aware manner, as illustrated in Figure 2b. On top of the learned embeddings, soft semantic labels are inferred via Lorentz Multinomial Logistic Regression (MLR), which maintains geometric consistency with the hyperbolic space and yields more suitable decision boundaries for curved latent structures. These labels further guide contrastive and consistency constraints, enabling the model to capture shared semantic information across multiple views.

Our main contributions are summarized as follows:

- **A new hyperbolic MVC framework:** We propose WAH-MVC to jointly preserve hierarchical structures, align global manifold distributions, and enhance feature discriminability on the Lorentz manifold of hyperbolic space.
- **A novel Lorentz alignment mechanism:** We introduce an Alignment mechanism based on the Lorentz SWD, enabling effective distribution-level alignment across views.
- **Experimental effectiveness:** Extensive empirical evaluations show the superiority of WAH-MVC over several SOTA methods in MVC.

Preliminary

In this section, we give a brief introduction to the Lorentz model of hyperbolic space and the Wasserstein Distance on Riemannian manifolds.

Hyperbolic Lorentz Model

The n-dimensional Lorentz model of hyperbolic space with constant negative curvature $K < 0$ is defined as

$$\mathbb{L}_K^n = \left\{ x \in \mathbb{R}^{n+1} \mid \langle x, x \rangle_{\mathcal{L}} = \frac{1}{K},\ x_0 > 0 \right\}, \quad (1)$$

where $\langle x, y \rangle_{\mathcal{L}} = -x_0 y_0 + \sum_{i=1}^{n} x_i y_i$ denotes the Lorentz inner product. The geodesic distance between two points is given by

$$d_{\mathbb{L}}(x, y) = \frac{1}{\sqrt{|K|}} \cdot \operatorname{arccosh}\left(|K| \langle x, y \rangle_{\mathcal{L}}\right), \quad (2)$$

where $x, y \in \mathbb{L}_K^n$. In our work, we adopt the canonical origin $\mathbf{x}_{\mathrm{o}} = [\sqrt{-1/K}, 0, \ldots, 0]$ as the reference point for tangent-space operations. At $\mathbf{x}_{\mathrm{o}}$, the tangent space $T_{\mathbf{x}_{\mathrm{o}}} \mathbb{L}_K^n$ is defined as the set of all vectors in $\mathbb{R}^{n+1}$ orthogonal to $\mathbf{x}_{\mathrm{o}}$ with respect to the Lorentz inner product. The exponential

map at $\mathbf{x}_o$, which projects a tangent vector $v \in T_{\mathbf{x}_o}\mathbb{L}_K^n$ onto the manifold, is defined as

$$\exp_{\mathbf{x}_o}^K(v) = \cosh(\alpha)\,\mathbf{x}_o + \sinh(\alpha)\,\frac{v}{\alpha}, \tag{3}$$

where $\alpha = \sqrt{|K|}\,\|v\|_{\mathcal{L}}$ and $\|v\|_{\mathcal{L}} = \sqrt{\langle v, v\rangle_{\mathcal{L}}}$ denotes the Lorentz norm. We present only the Lorentz operators utilized in this paper, while other operators and their detailed formulations are provided in the Appendix A.3.2.

Wasserstein Distance on Riemannian Manifolds

Wasserstein distance serves as the core cost metric in OT, and has been successfully applied in various tasks involving distribution alignment and semantic matching. Let $\mu, \nu \in \mathcal{P}_p(\mathcal{M})$ be two probability measures defined on a Riemannian manifold $\mathcal{M}$. The p-Wasserstein distance between μ and ν is given by

$$\mathcal{W}_p(\mu,\nu) = \min_{\gamma\in\Pi(\mu,\nu)} \int_{\mathcal{M}\times\mathcal{M}} d_{\mathcal{M}}(\mathbf{x},\mathbf{y})^p\, d\gamma(\mathbf{x},\mathbf{y}), \tag{4}$$

where $d_{\mathcal{M}}$ denotes the geodesic distance on $\mathcal{M}$, and $\Pi(\mu,\nu)$ is the set of all couplings (*i.e.*, joint distributions) with marginals μ and ν. In our setting, we treat hyperbolic embeddings $\mathcal{Z}_{\text{hyp}} = \{\mathbf{z}_i\}_{i=1}^N$ on the manifold as an empirical measure $\mu = \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{z}_i}$, where $\delta_{\mathbf{z}}$ denotes the Dirac measure centered at $\mathbf{z}$. Based on this construction, feature alignment across views reduces to computing the Wasserstein distance between empirical distributions on the manifold.

Method

As shown in Figure 3, the proposed WAH-MVC framework comprises three main components: Lorentz feature embedding, Wasserstein-Alignment for feature distribution, and contrastive clustering enhancement. The following subsections provide a detailed description.

Lorentz Feature Embedding

Given an M-view dataset $\mathcal{X} = \{X_1, \ldots, X_M\}$, where $X_m = \{\mathbf{x}_i^m \in \mathbb{R}^{D_m}\}_{i=1}^N$, N is the number of samples and D_m the feature dimension of view m. Our goal is to learn representations capturing both view-specific and shared information for clustering semantically consistent samples.

For each view m, we employ a geometry-aware encoder that maps the input features onto the Lorentz manifold. Specifically, a Euclidean encoder extracts latent features from the input X_m, which are then mapped to the $(d+1)$-dimensional Lorentz manifold via the exponential map $\exp_{\mathbf{x}_o}(\cdot)$ at the Lorentz origin $\mathbf{x}_o$. To further reduce dimensionality and enhance discriminability,we apply a Lorentz Fully Connected (FC) layer with curvature-aware normalization (Bdeir, Schwethelm, and Landwehr 2024; Chen et al. 2025a,b) within the Lorentz space. The overall transformation can be written as:

$$\tilde{\mathcal{Z}}_{\text{hyp}}^{(m)} = \text{LorentzFC}\left(\exp_{\mathbf{x}_o}(f_{\text{enc}}(X_m;\theta_{\text{enc}}^m))\right), \tag{5}$$

where $f_{\text{enc}}(\cdot)$ is the Euclidean encoder with view-specific parameters θ_{enc}^m. The Lorentz FC is a generalization of the Euclidean FC layer to Lorentzian geometry (detailed in App. A.3.2). This process yields the final hyperbolic embeddings, denoted by $\tilde{\mathcal{Z}}_{\text{hyp}}^{(m)} = \{\tilde{\mathbf{z}}_i^m \in \mathbb{L}_K^r\}_{i=1}^N$.

Wasserstein-Alignment for Feature Distribution

To enhance semantic consistency across views in hyperbolic space, we propose a global Wasserstein-Alignment strategy that operates directly on the Lorentz manifold. However, computing Wasserstein distance using Eq. 4 in curved spaces such as hyperbolic spaces poses significant computational challenges due to the nonlinearity of geodesic distances and the complexity of coupling space. As a countermeasure, we adopt a scalable approximation based on SWD. Specifically, following Bonet et al. (Bonet et al. 2023), we leverage two hyperbolic extensions of Horospherical Hyperbolic Sliced-Wasserstein (HHSW) and Geodesic Hyperbolic Sliced-Wasserstein (GHSW)—both of which are formulated within the Lorentz model. In this work, we focus on the HHSW variant due to its favorable computational properties and compatibility with our framework.

Formally, for probability measures $\mu, \nu \in \mathcal{P}_p(\mathbb{L}_K^n)$ on the Lorentz model, the HHSW distance is defined as

$$\text{HHSW}_\theta^p(\mu,\nu) = \int_{T_{\mathbf{x}_o}\mathbb{L}_K^n \cap \mathbb{S}^{n-1}} \mathcal{W}_p(B_\theta\#\mu, B_\theta\#\nu)\, d\lambda(\theta), \tag{6}$$

where $\mathbf{x}_o$ is the lorentz origin, and $\mathbb{S}^{n-1}$ is the unit sphere in the tangent space $T_{\mathbf{x}_o}\mathbb{L}_K^n$. The operator B_θ denotes the Busemann projection along the horospherical direction θ, which maps points in $\mathbb{L}_K^n$ to scalar values in R. The symbol $\mathcal{W}_p$ represents the standard p-Wasserstein distance in the real line. Since the projected distributions $B_\theta\#\mu$ and $B_\theta\#\nu$ are one-dimensional, $\mathcal{W}_p$ can be computed efficiently using inverse cumulative distribution functions:

$$\mathcal{W}_p(B_\theta\#\mu, B_\theta\#\nu) = \int_0^1 \left|F^{-1}_{B_\theta\#\mu}(u) - F^{-1}_{B_\theta\#\nu}(u)\right|^p du, \tag{7}$$

with F^{-1} denoting the quantile function. The integration over directions θ is taken for the uniform measure λ on the unit sphere. This sliced formulation preserves the curvature of the hyperbolic space while greatly reducing computational complexity. As a result, it enables efficient and scalable alignment of global distributions across views, thus being well-suited for high-dimensional MVC tasks.

To implement HHSW in our multi-view setting, we first sample projection directions from $T_{\mathbf{x}_o}\mathbb{L}_K^n$. Given a batch of features from the m-th view $\tilde{\mathcal{Z}}_{hyp}^{(m)}$, we generate a set of L directions $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_\ell\}_{\ell=1}^L$, where θ_ℓ satisfies the Lorentz orthogonality constraint, *i.e.*, $\langle\theta_\ell,\theta_\ell\rangle_{\mathbb{L}} = 1, \langle\theta_\ell,\mathbf{x}_o\rangle_{\mathbb{L}} = 0$. This ensures that the directions lie on the unit sphere and respect the geometry of the data manifold.

Then, each feature point $\tilde{z}_i^m \in \tilde{\mathcal{Z}}_{hyp}^{(m)}$ is projected onto the real line via a Buseman function, yielding a scalar that reflects its position along the geodesic defined by direction θ_ℓ. The resulting set forms a one-dimensional distribution $B_{\theta_\ell}^m$ for each view. The Buseman function is detailed in App. A.4.1.

Whereafter, we compute the HHSW distance between every pair of views $(m,n) \in \mathcal{V}$ across all sampled directions to enforce multi-view alignment:

$$\text{HHSW}_{\theta_\ell}^p\left(\tilde{\mathcal{Z}}_{hyp}^{(m)}, \tilde{\mathcal{Z}}_{hyp}^{(n)}\right) = \mathcal{W}_p\left(B_{\theta_\ell}^m, B_{\theta_\ell}^n\right). \tag{8}$$

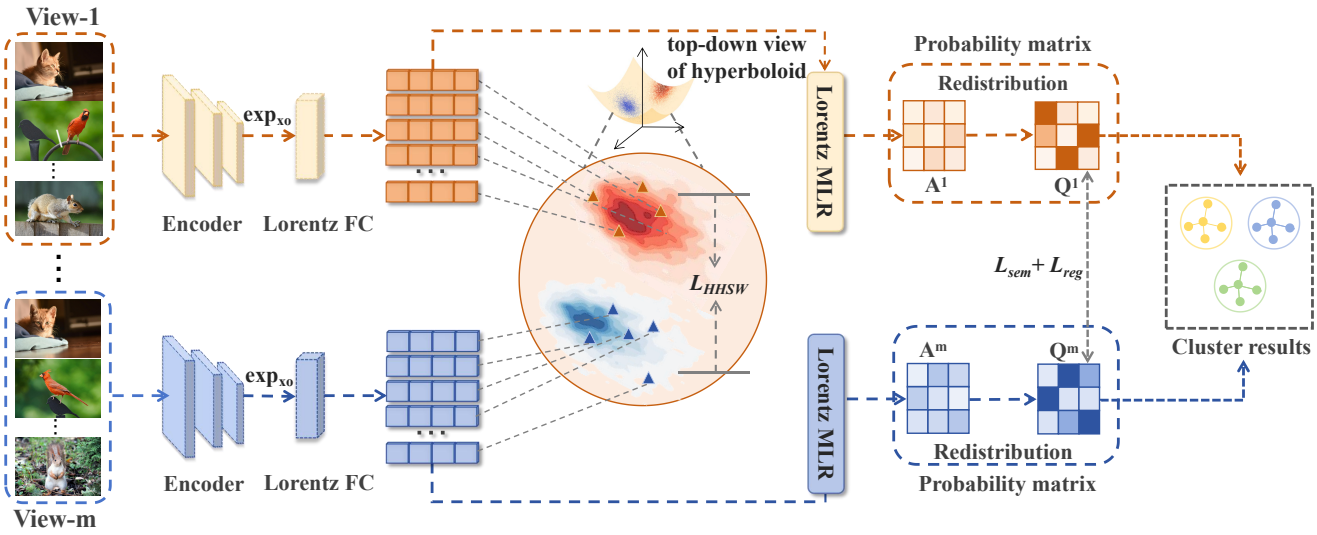


Figure 3: The framework of WAH-MVC. It consists of three main components: (1) **Lorentz Feature Embedding**, which maps view-specific features into a curvature-aware Lorentz manifold for improved representation learning; (2) **Wasserstein-Alignment for Feature Distribution** ($\mathcal{L}_{\text{HHSW}}$), which aligns distributions of different views in hyperbolic space using a sliced Wasserstein distance; (3) **Contrastive Cluster Enhancement** ($\mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{reg}}$), which enhances cluster discriminability by enforcing semantic consistency across views through contrastive learning.

Finally, the HHSW-based alignment loss is defined as the average discrepancy across all view pairs and projection directions, as illustrated below:

$$\mathcal{L}_{\text{HHSW}} = \frac{1}{|\mathcal{V}| \times L} \sum_{(m,n)\in\mathcal{V}} \sum_{\ell=1}^{L} \text{HHSW}_{\theta_\ell}^{p}\left(\tilde{\mathcal{Z}}_{hyp}^{(m)}, \tilde{\mathcal{Z}}_{hyp}^{(n)}\right), \quad (9)$$

where $\mathcal{V}$ is the set of unordered view pairs. This loss function provides a geometry-aware and computationally efficient mechanism for aligning multi-view representations on the Lorentz manifold.

Contrastive Cluster Enhancement

To improve clustering discriminability, we introduce a contrastive learning strategy that exploits cross-view semantic consistency. After aligning the Lorentz representations via the SWD-based method, we employ Lorentz MLR (Bdeir, Schwethelm, and Landwehr 2024) on the Lorentz embeddings to generate class probabilities. As a Lorentz extension of the Euclidean MLR, Lorentz MLR performs classification by defining distance-based decision boundaries that respect the curvature of the Lorentz manifold. The cluster probability matrix is computed by

$$\mathcal{A}^{(m)} = \text{LorentzMLR}(\tilde{\mathcal{Z}}_{\text{hyp}}^{(m)}) = \{\mathbf{a}_i^m \in \mathbb{R}^K\}_{i=1}^N, \quad (10)$$

where K represents the number of clusters. The detailed formulation of LorentzMLR is provided in App. A.3.2.

To refine cluster assignments, we compute a target distribution $\mathbf{Q}^{(m)} \in \mathbb{R}^{N\times K}$ from $\mathcal{A}^{(m)}$, following the method proposed in (Cui et al. 2023):

$$q_{ij}^{(m)} = \frac{(a_{ij}^{(m)})^2 / \sum_{i=1}^{N} a_{ij}^{(m)}}{\sum_{k=1}^{K}\left[(a_{ik}^{(m)})^2 / \sum_{i=1}^{N} a_{ik}^{(m)}\right]}, \quad (11)$$

Let $\mathbf{q}_k^{(m)}$ denote the k-th column of the matrix $\mathbf{Q}^{(m)}$, where $\mathbf{q}_k^{(m)} = [q_{1k}^{(m)}, q_{2k}^{(m)}, \ldots, q_{Nk}^{(m)}]^\top$. Here, $q_{ij}^{(m)}$ denotes the probability of assigning sample i to cluster j in view m. This weighting mechanism boosts confident predictions while downplays ambiguous ones, thereby enabling more discriminative clustering.

Subsequently, the similarity between cluster-wise soft assignments is measured by the inner product, shown below:

$$s_{k,k}^{(m,n)} = \left(\mathbf{q}_k^{(m)}\right)^\top \mathbf{q}_k^{(n)}. \quad (12)$$

Based on these cross-view similarities, a contrastive loss that encourages alignment between matched clusters while distinguishing mismatched ones is constructed. Inspired by (Chen et al. 2023), the contrastive loss between views m and n is formulated as:

$$\mathcal{L}_{\text{c}} = -\frac{1}{K}\sum_{k=1}^{K} \log \frac{\exp\left(s_{k,k}^{(m,n)}/\tau\right)}{\exp\left(s_{k,k}^{(m,n)}/\tau\right) + \sum_{j=1,j\neq k}^{K} \exp\left(s_{k,j}^{(m,n)}/\tau\right)}, \quad (13)$$

with τ denoting a temperature parameter controlling the sharpness of the similarity distribution. Thereby, the total contrastive loss across all views is expressed as:

$$\mathcal{L}_{\text{sem}} = \sum_{m=1}^{M} \sum_{n=1,\, n\neq m}^{M} \mathcal{L}_{\text{c}}(m,n). \quad (14)$$

To avoid degenerate assignments where all instances collapse into a single cluster, we follow (Cui et al. 2023) and incorporate a cross-view regularization term as follows:

$$\mathcal{L}_{\text{reg}} = \sum_{m=1}^{M} \sum_{j=1}^{K} p_j^{(m)} \log p_j^{(m)}, \quad (15)$$

where $p_j^{(m)} = \frac{1}{N}\sum_{i=1}^{N} q_{ij}^{(m)}$ is an entropy-based regularizer that encourages balanced cluster assignments across views, thus enhancing cross-view consistency.

Optimization and Label Inference

Previously, we have detailed the global alignment strategy via SWD and the semantic-aware contrastive learning module that captures shared semantics across views. Here, we summarize the overall training objective of WAH-MVC, which integrates three complementary loss functions:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{HHSW}} + \beta\mathcal{L}\text{sem} + \gamma\mathcal{L}\text{reg}, \tag{16}$$

where α, β, and γ are three hyperparameters used to balance the contribution of each component.

After training, the final semantic label for the i-th ($i \in 1, 2, \ldots, N$) instance is predicted by averaging its soft assignments across all M views and selecting the cluster with the highest mean probability, as shown below:

$$y_i = \arg\max_j \left(\frac{1}{M} \sum_{m=1}^{M} q_{ij}^{(m)} \right). \tag{17}$$

This simple yet effective voting mechanism ensures robust prediction by leveraging consensus across views.

Experiments

Experimental Settings

Datasets. We evaluate WAH-MVC on six publicly available multi-view datasets. MNIST-USPS (Peng et al. 2019) is a two-view dataset, with each view containing 5,000 handwritten digit images from different domains. COIL-10 (Xu et al. 2021) consists of 720 grayscale images from 10 categories, with each object captured from three different poses. Fashion (Xiao, Rasul, and Vollgraf 2017) includes 10,000 samples from 10 clothing categories, each observed from front, side, and back views. Scene-15 (Fei-Fei and Perona 2005) comprises 4,485 images from 15 classes, each represented by three types of handcrafted features. Amazon (Saenko et al. 2010) contains 4,790 color images from 10 categories under four views. Youtube Video (Madani, Georg, and Ross 2012) is a large-scale dataset with 101,499 samples from 31 classes, each represented by three feature vectors of dimensions 512, 647, and 838.

Comparative Methods. The following methods are selected for comparison: view-level contrastive methods, such as DSIMVC (Tang and Liu 2022a), CPSPAN (Jin et al. 2023), MFLVC (Xu et al. 2022b), and DCMVC (Cui et al. 2024), aim to maximize feature consistency across views; clustering-level contrastive methods, including CVCL (Chen et al. 2023) and SCM (Luo et al. 2024), focusing on enhancing clustering structure through assignment-level contrast; aggregation-based methods, such as DSMVC (Tang and Liu 2022b), GCFAgg (Yan et al. 2023), and MVCAN (Xu et al. 2024), integrate multi-view features with adaptive or global strategies. We also include an OT-based method, CSOT (Zhang et al. 2024), improving MVC performance via Euclidean-based semantic alignment.

Network Architecture and Parameter Settings. The proposed model employs a hybrid Euclidean encoder $f_{\text{enc}}(\cdot)$ to process both vector and image inputs. For the m-th vector view, features are encoded using a ReLU-activated MLP with 3–5 hidden layers. A typical 3-layer configuration is $[D_m, 256, 512, d]$, where D_m denotes the input dimension and d represents the shared output size. For the m-th image view, a lightweight CNN is employed, consisting of two convolutional blocks, an adaptive pooling layer, and 2–3 fully connected layers. The final output is a feature vector with dimension d, selected from $\{256,\ 512,\ 1024\}$. The features are mapped onto the Lorentz manifold via $\exp_{\mathbf{x}_o}(\cdot)$ and transformed by a LorentzFC layer to produce the final hyperbolic embeddings (see App. A.5.2 for details). The total loss is weighted by hyperparameters α, β, and γ, selected using grid search from the set $\{0.001, 0.005, 0.01, 0.05, 0.1, 1.0\}$. A temperature parameter τ controls the sharpness of the contrastive objective. Its selection process is detailed in the following Ablation Studies section.

Performance Evaluation

Table 1 summarizes the clustering performance of different methods on six benchmark datasets, with the best results highlighted in bold. Overall, WAH-MVC consistently surpasses the baselines, demonstrating strong effectiveness in multi-view clustering.

On relatively simple and balanced datasets (MNIST-USPS, COIL-10, Fashion), the most recent methods perform competitively. Nonetheless, WAH-MVC consistently outperforms strong baselines (e.g., CVCL, SCM, CSOT) across all metrics. On more challenging datasets such as Scene-15 and Amazon, WAH-MVC achieves substantial improvements, surpassing the second-best methods by approximately 29.8% (ACC) and 32.1% (NMI) on Scene-15. Despite the inherent challenges of the YoutubeVideo dataset, including substantial inter-view heterogeneity, label noise, and large-scale complexity, WAH-MVC achieves the highest ACC among all compared methods. Specifically, it outperforms the previous SOTA method GCFAgg by 1.82% in ACC, showcasing its good generalization ability. Although WAH-MVC reports lower NMI than GCFAgg, this may be attributed to NMI's sensitivity to fragmented or noisy labels (Vinh, Epps, and Bailey 2010). In contrast, the ACC metric captures overall assignment accuracy and may better reflect WAH-MVC's ability to preserve dominant semantics under noisy ground truth.

These results collectively highlight two major advantages of WAH-MVC. Firstly, unlike contrastive learning methods such as CVCL and DSMVC that operate in Euclidean space with limited ability to model the intrinsic hierarchical structures of multi-view data, WAH-MVC leverages hyperbolic space to preserve hierarchy and effectively capture multi-level dependencies among views and clusters. Secondly, the OT-based distribution alignment on the Lorentz manifold allows WAH-MVC to mitigate view discrepancies at the semantic level, thereby facilitating learning a more discriminative hyperbolic network embedding.

Method	MNIST-USPS (V=2, N=5000)		COIL-10 (V=3, N=720)		Scene-15 (V=3, N=4485)		Amazon (V=4,N=4790)		Fashion (V=3, N=10000)		YoutubeVideo (V=3, N=101499)	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
DSIMVC (ICML'22)	99.34	98.13	99.68	99.35	28.27	29.04	64.80	62.46	95.50	92.17	15.57	7.46
MFLVC (CVPR'22)	99.56	98.73	92.50	92.76	34.94	29.10	89.44	91.19	99.25	98.11	21.55	22.88
DSMVC (CVPR'22)	96.34	94.27	96.39	95.32	43.48	41.11	38.64	28.29	79.40	77.98	12.47	10.35
CVCL (ICCV'23)	99.70	99.13	99.43	99.04	44.59	42.17	83.10	67.49	99.31	99.21	22.61	22.46
CPSPAN (CVPR'23)	93.30	87.61	82.92	89.80	22.92	15.69	59.90	53.30	72.22	75.67	23.88	22.24
GCFAgg (CVPR'23)	99.56	98.71	85.83	94.08	42.27	42.56	90.58	86.23	98.97	97.38	27.58	**27.31**
MVCAN (CVPR'24)	99.24	98.01	99.31	98.57	38.32	39.48	83.15	85.28	82.89	86.14	25.01	24.43
DCMVC (TIP'24)	95.76	90.27	85.83	95.15	32.60	28.34	68.46	62.43	94.89	92.42	27.26	27.29
SCM (IJCAI'24)	98.94	97.02	99.86	99.68	40.83	38.89	48.98	40.20	98.00	95.80	18.10	18.49
CSOT (TIP'24)	99.22	99.22	99.58	99.24	35.83	29.82	98.04	95.39	99.12	97.81	21.11	21.57
WAH-MVC	**99.88**	**99.62**	**99.91**	**99.74**	**74.38**	**74.68**	**99.88**	**99.62**	**99.74**	**99.29**	**29.40**	22.74

Table 1: Comparison of clustering performance on six multi-view datasets (V: Number of Views; N: Samples per View)

$\mathcal{L}_{\text{HHSW}}$	$\mathcal{L}_{\text{sem}}$	$\mathcal{L}_{\text{reg}}$	Amazon (V=4, N=4790)		Fashion (V=3, N=10000)		YoutubeVideo (V=3, N=101499)	
			ACC	NMI	ACC	NMI	ACC	NMI
✗	✓	✓	99.38	98.37	99.55	98.81	25.49	21.46
✓	✗	✓	27.57	20.23	27.28	15.53	13.26	3.41
✓	✓	✗	85.10	93.94	71.38	86.98	27.23	12.46
✓	✓	✓	**99.88**	**99.62**	**99.74**	**99.29**	**29.40**	**22.74**

Table 2: Impact of each loss component.

Method	Amazon		Fashion		YoutubeVideo	
	ACC	NMI	ACC	NMI	ACC	NMI
CVCL	83.10	67.49	99.31	98.21	22.61	22.46
CSOT	98.04	95.39	99.12	97.81	21.11	21.57
WAH-MVC w/ A	99.36	98.27	99.23	98.04	26.78	21.82
WAH-MVC w/ B	99.62	98.93	99.22	98.01	28.60	22.20

Table 3: Clustering performance with different WAH-MVC backbones. A and B denote the encoders from CVCL and CSOT, respectively.

Dataset	Method	ACC	NMI	Time (s)
Scene-15 (V=3, N=4485)	w/ $\mathcal{L}_{\text{HCL}}$	72.27	73.32	5.95
	w/ $\mathcal{L}_{\text{HHSW}}$	74.38	74.68	5.85
Amazon (V=4, N=4790)	w/ $\mathcal{L}_{\text{HCL}}$	99.53	98.52	2.77
	w/ $\mathcal{L}_{\text{HHSW}}$	99.78	99.39	2.67
YoutubeVideo (V=3, N=101499)	w/ $\mathcal{L}_{\text{HCL}}$	22.45	23.19	53.32
	w/ $\mathcal{L}_{\text{HHSW}}$	29.40	22.74	52.22

Table 4: Clustering performance and training time comparison of $\mathcal{L}_{\text{HCL}}$ and $\mathcal{L}_{\text{HHSW}}$ loss-based hyperbolic alignment methods. Time: per-epoch training time (in seconds).

Method	Fashion		Scene-15		Amazon	
	ACC	NMI	ACC	NMI	ACC	NMI
GHSW	99.31	98.23	72.77	73.47	99.77	99.33
HHSW	99.74	98.29	74.38	74.68	99.88	99.62

Table 5: Projection Comparison: HHSW vs. GHSW

Ablation Studies

In this part, we make a series of ablation studies to explore the impact of potential factors on the model performance.

Loss functions. Here, we conduct clustering experiments on the Amazon, Fashion, and YoutubeVideo datasets to analyze the significance of each loss function in WAH-MVC. As shown in Table 2, the full model that integrates all three losses consistently achieves the best performance across all datasets. Notably, removing $\mathcal{L}_{\text{sem}}$ results in a drastic drop in clustering quality, highlighting its necessity for semantic discriminability. The regularization term $\mathcal{L}_{\text{reg}}$ also enhances learning stability, especially on mid-scale datasets like Fashion. As the number of samples increases under a fixed number of views, the impact of $\mathcal{L}_{\text{HHSW}}$ becomes more pronounced. On the large-scale and highly heterogeneous YoutubeVideo dataset, incorporating $\mathcal{L}_{\text{HHSW}}$ respectively improves ACC and NMI by 3.91% and 1.28%, over the variant without using it. These findings demonstrate that $\mathcal{L}_{\text{HHSW}}$ effectively aligns multi-view distributions at the semantic level, thereby improving clustering performance.

Backbone models. To verify the robustness of WAH-MVC to different backbone architectures, we choose the encoders of CVCL and CSOT, which are used solely as feature extractors in our framework. As shown in Table 3, WAH-MVC consistently achieves high clustering ACC and NMI on the Amazon and Fashion datasets, regardless of the encoder employed. This demonstrates that our model is insensitive to backbone variations. Even on the challenging YoutubeVideo dataset, where overall performance is lower, WAH-MVC still delivers notable improvements over the corresponding baseline models under the same encoders. Therefore, we argue that the strength of WAH-MVC stems from its core hyperbolic alignment and clustering mechanisms, rather than from a specific encoder design.

Alignment mechanisms. Table 4 shows the effective-

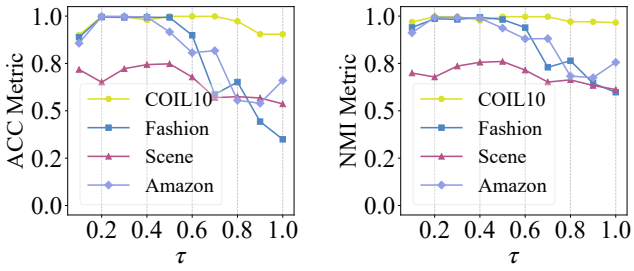


Figure 4: Effect of τ on ACC and NMI for five datasets.

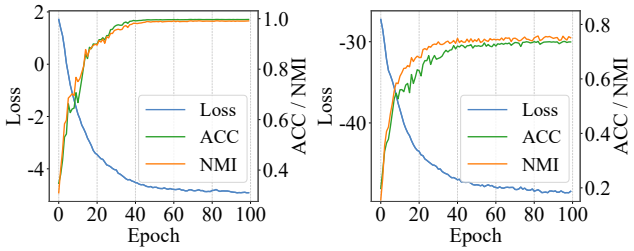
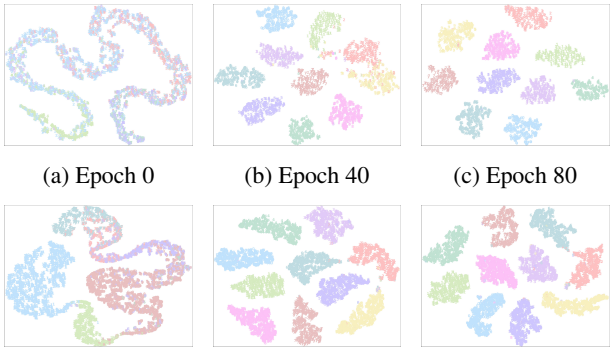


Figure 5: Convergence analysis of WAH-MVC on Amazon (left) and Scene-15 (right).

ness of $\mathcal{L}_{\text{HHSW}}$ across Scene-15, Amazon, and YoutubeVideo. Compared to the hyperbolic instance-level contrastive loss $\mathcal{L}_{\text{HCL}}$ (see App. A.5.2), our method consistently achieves superior or comparable results. (1) **Clustering performance:** $\mathcal{L}_{\text{HHSW}}$ outperforms $\mathcal{L}_{\text{HCL}}$ by 2.11% (ACC) and 1.36% (NMI) on Scene-15, and improves ACC by 6.95% on YoutubeVideo, highlighting its advantage in modeling global multi-view semantics. (2) **Complexity:** $\mathcal{L}_{\text{HCL}}$ requires $\mathcal{O}(M^2B^2r)$ time due to pairwise comparisons, while $\mathcal{L}_{\text{HHSW}}$ has a complexity of $\mathcal{O}(M^2LB(r + \log B))$ due to efficient sorting over L projections. On large datasets YoutubeVideo, $\mathcal{L}_{\text{HHSW}}$ achieves faster training than $\mathcal{L}_{\text{HCL}}$. Complexity details are provided in the App. A.5.3.

HHSW vs. GHSW. The key difference between HHSW and GHSW lies in their projection mappings (Bonet et al. 2023): HHSW uses a horospherical projection via the Busemann function, while GHSW employs a geodesic projection (see App. A.4 for details). We evaluate the impact of these two mapping strategies on the Fashion, Scene-15, and Amazon datasets using the Wasserstein distance with $p = 2$. As shown in Table 5, HHSW consistently outperforms GHSW across all evaluation metrics on all datasets. This demonstrates that the Busemann-based projection in HHSW better preserves semantic structures relevant to clustering.

Temperature parameter. In this part, we investigate the effect of the temperature parameter τ in Eq. 13 on the clustering performance of WAH-MVC. Figure 4 reports ACC and NMI on five different datasets as τ varies in the range $\{0.1, 1.0\}$. As shown, a moderate value range (*e.g.*, $\tau = [0.3, 0.5]$) generally yields better clustering performance, whereas excessively small or large values of τ deteriorate results. This is because τ balances sample discrimination hardness: too small a τ may over-emphasize hard negatives,



(a) Epoch 0 (b) Epoch 40 (c) Epoch 80

(d) Epoch 0 (e) Epoch 40 (f) Epoch 80

Figure 6: 2D visualization of the learned features at different epochs. (a–c): Amazon; (d–f): Fashion.

while too large a τ will over-smooth the similarity distribution. Therefore, choosing a proper τ improves both the stability and effectiveness of the contrastive learning process.

Visualization. Figure 5 plots the ACC, NMI, and total loss in Eq. 16 over training epochs on the Amazon and Scene-15 datasets. The total loss steadily decreases and quickly plateaus, while the ACC and NMI rise consistently before stabilizing, indicating that WAH-MVC converges reliably and maintains training stability. Moreover, Figure 6 visualizes the evolution of the learned features across epochs. This is realized using the t-SNE technique (Maaten and Hinton 2008), which projects high-dimensional features into a 2D space. As training progresses, intra-cluster compactness increases and inter-cluster ambiguity decreases, showing that WAH-MVC learns more discriminative multi-view representations despite semantic gaps.

Conclusion

In this paper, we propose WAH-MVC, a novel hyperbolic multi-view clustering framework that learns view-invariant representations by aligning cross-view semantic structures. To this end, we first design a view-specific hyperbolic encoder to capture the latent hierarchical features within each view. At its core, WAH-MVC performs SWD-based cluster-level alignment strategy on the Lorentz manifold, explicitly capturing the shared semantics among views and driving the model toward more discriminative and consistent cluster assignments. Extensive experiments and ablation studies on several benchmarking datasets demonstrate the superior performance of WAH-MVC and quantify the contribution of each component to the overall objective.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (62306127, 62020106012, 62332008), the Natural Science Foundation of Jiangsu Province (BK20231040), the Fundamental Research Funds for the Central Universities (JUSRP124015), the Key Project of Wuxi Municipal Health Commission (Z202318), the EU Horizon project ELIAS (101120237), the FIS project

GUIDANCE (FIS2023-03251), and the National Key R&D Program of China (2023YFF1105102, 2023YFF1105105).

References

Bdeir, A.; Schwethelm, K.; and Landwehr, N. 2024. Fully Hyperbolic Convolutional Neural Networks for Computer Vision. In *International Conference on Learning Representations*, 47687–47711.

Bonet, C.; Chapel, L.; Drumetz, L.; and Courty, N. 2023. Hyperbolic Sliced-Wasserstein via Geodesic and Horospherical Projections. In *Proceedings of the Annual Workshop on Topology, Algebra, and Geometry in Machine Learning*, 334–370.

Chami, I.; Wolf, A.; Juan, D.-C.; Sala, F.; Ravi, S.; and Ré, C. 2020. Low-Dimensional Hyperbolic Knowledge Graph Embeddings. arXiv:2005.00545.

Chen, J.; Mao, H.; Woo, W. L.; and Peng, X. 2023. Deep Multiview Clustering by Contrasting Cluster Assignments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 16752–16761.

Chen, J.; Yang, S.; Mao, H.; and Fahy, C. 2022. Multiview Subspace Clustering Using Low-Rank Representation. *IEEE Transactions on Cybernetics*, 52(11): 12364–12378.

Chen, J.; Yang, S.; Peng, X.; Peng, D.; and Wang, Z. 2024. Augmented Sparse Representation for Incomplete Multiview Clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 35(3): 4058–4071.

Chen, Z.; Song, Y.; Wu, X.; and Sebe, N. 2025a. Gyrogroup Batch Normalization. In *The Thirteenth International Conference on Learning Representations*.

Chen, Z.; Wu, X.-J.; Schölkopf, B.; and Sebe, N. 2025b. Riemannian Batch Normalization: A Gyro Approach. *arXiv:2509.07115*.

Cui, C.; Ren, Y.; Pu, J.; Pu, X.; and He, L. 2023. Deep Multi-View Subspace Clustering with Anchor Graph. *arXiv:2305.06939*.

Cui, J.; Li, Y.; Huang, H.; and Wen, J. 2024. Dual Contrast-Driven Deep Multi-View Clustering. *IEEE Transactions on Image Processing*, 33: 4753–4764.

Fei-Fei, L.; and Perona, P. 2005. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 524–531.

Guo, K.; Zhao, C.; and Wang, J. 2024. A fast mask synthesis method for face recognition. *Visual Intelligence*, 2(25).

Guo, R.; Yang, M.; Lin, Y.; Peng, X.; and Hu, P. 2024. Robust Contrastive Multi-view Clustering against Dual Noisy Correspondence. *Advances in Neural Information Processing Systems*, 37.

Hu, M.; and Chen, S. 2019. One-Pass Incomplete Multi-View Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3838–3845.

Hu, S.; Tian, B.; Liu, W.; and Ye, Y. 2025. Self-supervised Trusted Contrastive Multi-view Clustering with Uncertainty Refined. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 17305–17313.

Jin, J.; Wang, S.; Dong, Z.; Liu, X.; and Zhu, E. 2023. Deep Incomplete Multi-View Clustering With Cross-View Partial Sample and Prototype Alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11600–11609.

Li, L.; Wan, Z.; and He, H. 2023. Incomplete Multi-View Clustering With Joint Partition and Graph Learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(1): 589–602.

Li, Y.; Yang, M.; Peng, D.; Li, T.; Huang, J.; and Peng, X. 2022. Twin Contrastive Learning for Online Clustering. *International Journal of Computer Vision*, 130(9): 2205–2221.

Li, Z.; Wang, Q.; Tao, Z.; Gao, Q.; Yang, Z.; et al. 2019. Deep Adversarial Multi-view Clustering Network. In *International Joint Conference on Artificial Intelligence*, 2952–2958.

Lin, F.; Bai, B.; Bai, K.; Ren, Y.; Zhao, P.; and Xu, Z. 2022. Contrastive Multi-view Hyperbolic Hierarchical Clustering. arXiv:2205.02618.

Lin, F.; Bai, B.; Guo, Y.; Chen, H.; Ren, Y.; and Xu, Z. 2023. MHCN: A Hyperbolic Neural Network Model for Multi-view Hierarchical Clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 16525–16535.

Liu, X.; Li, M.; Tang, C.; Xia, J.; Xiong, J.; Liu, L.; Kloft, M.; and Zhu, E. 2021. Efficient and Effective Regularized Incomplete Multi-View Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8): 2634–2646.

Liu, X.; Zhu, X.; Li, M.; Wang, L.; Tang, C.; Yin, J.; Shen, D.; Wang, H.; and Gao, W. 2019. Late Fusion Incomplete Multi-View Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10): 2410–2423.

Luo, C.; Xu, J.; Ren, Y.; Ma, J.; and Zhu, X. 2024. Simple Contrastive Multi-View Clustering with Data-Level Fusion. In *International Joint Conference on Artificial Intelligence*, 4697–4705.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605.

Madani, O.; Georg, M.; and Ross, D. A. 2012. On Using Nearly-Independent Feature Families for High Precision and Confidence. In *Proceedings of the Asian Conference on Machine Learning*, 269–284.

Nickel, M.; and Kiela, D. 2018. Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry. In *International Conference on Machine Learning*, 3779–3788.

Peng, X.; Huang, Z.; Lv, J.; Zhu, H.; and Zhou, J. T. 2019. COMIC: Multi-view Clustering Without Parameter Selection. In *International Conference on Machine Learning*, 5092–5101.

Peng, X.; Li, Y.; Tsang, I. W.; Zhu, H.; Lv, J.; and Zhou, J. T. 2022. XAI Beyond Classification: Interpretable Neural Clustering. *Journal of Machine Learning Research*, 23(6): 1–28.

Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting Visual Category Models to New Domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 213–226.

Tang, H.; and Liu, Y. 2022a. Deep Safe Incomplete Multi-view Clustering: Theorem and Algorithm. *Advances in Neural Information Processing Systems*, 162.

Tang, H.; and Liu, Y. 2022b. Deep Safe Multi-View Clustering: Reducing the Risk of Clustering Performance Degradation Caused by View Increase. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 202–211.

Tao, Z.; Li, J.; Fu, H.; Kong, Y.; and Fu, Y. 2021. From Ensemble Clustering to Subspace Clustering: Cluster Structure Encoding. *IEEE Transactions on Neural Networks and Learning Systems*, 34(5): 2670–2681.

Vinh, N. X.; Epps, J.; and Bailey, J. 2010. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Advances in Neural Information Processing Systems*, 11.

Wang, Q.; Tao, Z.; Xia, W.; Gao, Q.; Cao, X.; and Jiao, L. 2023. Adversarial Multiview Clustering Networks With Adaptive Fusion. *IEEE Transactions on Neural Networks and Learning Systems*, 34(10): 7635–7647.

Wei, Z.; Xu, C.; Guan, Z.; and Liu, Y. 2020. Multiview Concept Learning via Deep Matrix Factorization. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2): 814–825.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv:1708.07747*.

Xiao, J.; Fan, Y.; Sun, R.; Wang, J.; and Luo, Z.-Q. 2022. Stability Analysis and Generalization Bounds of Adversarial Training. In *Advances in Neural Information Processing Systems*, 15446–15459.

Xie, X.; Wu, J.; Liu, G.; and Lin, Z. 2024. SSCNet: learning-based subspace clustering. *Visual Intelligence*, 2(11).

Xing, Y.; Song, Q.; and Cheng, G. 2021. On the Algorithmic Stability of Adversarial Training. *Advances in Neural Information Processing Systems*, 34.

Xu, J.; Ren, Y.; Tang, H.; Pu, X.; Zhu, X.; Zeng, M.; and He, L. 2021. Multi-VAE: Learning Disentangled View-Common and View-Peculiar Visual Representations for Multi-View Clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9234–9243.

Xu, J.; Ren, Y.; Tang, H.; Yang, Z.; Pan, L.; Yang, Y.; Pu, X.; Yu, P. S.; and He, L. 2022a. Self-Supervised Discriminative Feature Learning for Deep Multi-View Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(7): 7470–7482.

Xu, J.; Ren, Y.; Wang, X.; Feng, L.; Zhang, Z.; Niu, G.; and Zhu, X. 2024. Investigating and Mitigating the Side Effects of Noisy Views for Self-Supervised Clustering Algorithms in Practical Multi-View Scenarios. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 22957–22966.

Xu, J.; Tang, H.; Ren, Y.; Peng, L.; Zhu, X.; and He, L. 2022b. Multi-Level Feature Learning for Contrastive Multi-View Clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 16051–16060.

Yan, W.; Zhang, Y.; Lv, C.; Tang, C.; Yue, G.; Liao, L.; and Lin, W. 2023. GCFAgg: Global and Cross-View Feature Aggregation for Multi-View Clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 19863–19872.

Zhang, Q.; Zhang, L.; Song, R.; Cong, R.; Liu, Y.; and Zhang, W. 2024. Learning Common Semantics via Optimal Transport for Contrastive Multi-View Clustering. *IEEE Transactions on Image Processing*, 33: 4501–4515.

Zhang, Y.; Lin, Y.; Yan, W.; Yao, L.; Wan, X.; Li, G.; Zhang, C.; Ke, G.; and Xu, J. 2025. Incomplete Multi-view Clustering via Diffusion Contrastive Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 22650–22658.