Supplementary Materials for: Robust Multi-view Clustering via Pseudo Label Guided universum Learning

1 Introduction

This supplementary material provides further insights into the PAUSE framework to enhance the understanding of its design and functionality. In Section 2, the datasets, baseline methods, clustering techniques, and mixing strategy employed in our experiments are introduced. Section 3 offers a rigorous justification for the selection of K-means as the pseudo-labeling method. In Section 4, an additional detailed parameter analysis is provided, and Section 5 includes an ablation study focusing on the warm-up stage. Section 6 discusses the robustness analysis, and Section 7 presents further visualization results that illustrate the efficacy of the clustering outcomes. Finally, Section 8 concludes with a deeper theoretical analysis of the proposed UniLoss.

2 Datasets, Baselines, and Strategies

2.1 Datasets

In this section, we briefly introduce the benchmark multi-view datasets employed in our experiments.

CUB [17] comprises 600 images from 10 bird categories, with each image accompanied by both visual features and textual descriptions. This subset is derived from the Caltech-UCSD Birds-200-2011 dataset by selecting 10 representative bird categories and randomly sampling 600 images, following the approach in [6].

WIKI [14] is a benchmark dataset comprising 2,866 image–text pairs extracted from the featured articles of Wikipedia. While Wikipedia organizes its content into 29 thematic categories, only the 10 largest categories are retained. Each sample includes two views: an image view representing visual features and a text view providing the textual information.

MNIST-USPS is constructed following [13] by treating the USPS and MNIST datasets as two distinct views. For each dataset, 5,000 samples covering 10 digit classes are randomly selected. MNIST images are represented as 784-dimensional vectors, whereas USPS images are represented as 256-dimensional vectors.

NUS-WIDE [4] consists of 9,000 images paired with their corresponding captions from 10 classes. Visual features are extracted using a VGG19 network, and textual features are obtained via a Sentence CNN, as described in [29].

NoisyMNIST [18] is a multi-view version of the MNIST dataset. The original MNIST images serve as view 1, while view 2 is generated by randomly selecting images from the same classes and adding Gaussian noise. Although the full dataset contains 70,000 instances across 10 classes, we follow [10] and randomly select 30,000 instances for evaluation, as many baseline methods cannot handle datasets of such large scale.

2.2 Baselines

To validate the effectiveness and robustness of our proposed PAUSE, we compare it with several state-of-the-art multi-view clustering methods. A brief introduction to each baseline is provided below.

HCN [20] is an unsupervised multi-view clustering method that leverages canonical correlation analysis and contrastive learning to capture hierarchical consensus across views. It computes class-level, instance-level, and global consensus measures to integrate complementary information and yield more discriminative features.

MFLVC [21] is a multi-level feature learning framework for contrastive multi-view clustering. It extracts low-level, high-level, and semantic features in a fusion-free manner by enforcing a reconstruction objective on low-level features and contrastive consistency objectives on high-level features and semantic labels, thus mitigating the influence of view-specific noise.

CVCL [3] is a cross-view contrastive learning method that first extracts view-dependent features using deep autoencoders and then aligns cluster assignments across views via a cluster-level contrastive strategy, resulting in more robust clustering.

DealMVC [25] is a dual contrastive calibration network for multi-view clustering. It fuses cross-view features into a global representation and applies both global and local contrastive calibration losses to align view similarity graphs with high-confidence pseudo-labels, thereby regularizing the feature structure.

ProImp [9] is designed for incomplete multi-view clustering. It adopts a dual-stream model with a dual attention mechanism and dual contrastive losses to learn view-specific prototypes and model sample–prototype relationships. When a view is missing, the method recovers the missing data using the learned prototypes, preserving both instance commonality and view versatility.

MVCLN [24] is a noise-robust contrastive learning framework that simultaneously learns representations and establishes cross-view correspondences. It constructs positive pairs from known correspondences and negative pairs via random sampling, employing a noise-robust contrastive loss to mitigate the adverse impact of false negatives.

SURE [23] is a unified contrastive learning approach that addresses partial view unalignment and sample missing issues. It treats available cross-view pairs as positives and randomly samples negatives, with a noise-robust contrastive loss to reduce the influence of false negatives, thereby effectively handling incomplete information.

DCP [10] is a unified framework for incomplete multi-view representation learning that jointly optimizes cross-view consistency and missing view recovery. It maximizes mutual information across views via contrastive learning and minimizes conditional entropy for data recovery, achieving a provably minimal and sufficient representation.

GCFAgg [22] is a global and cross-view feature aggregation network for multi-view clustering. It aggregates features across samples and views to obtain a consensus representation, which is then aligned with view-specific representations via a structure-guided contrastive module to capture complementary information.

DIVIDE [11] is a multi-view clustering method that addresses false negatives and false positives by leveraging random walks to progressively identify globally consistent data pairs. It decouples inter-view and intra-view contrastive learning in separate embedding spaces, thereby enhancing clustering robustness and performance.

Candy [5] is a multi-view clustering approach that tackles the dual noisy correspondence problem by exploiting inter-view similarities as contextual cues and employing a spectral-based denoising module to refine positive and negative pair constructions, thus mitigating the adverse effects of false correspondences.

2.3 Clustering Methods

To validate the appropriateness of using K-means for pseudo-label generation, we compare it with several widely used clustering methods. A brief overview of these alternative approaches is provided below.

K-means [12] is a partition-based clustering algorithm that aims to minimize the within-cluster sum of squares. It iteratively assigns each data point to the nearest cluster centroid and recalculates the centroids until convergence, resulting in compact and well-separated clusters.

Agglomerative Clustering [19] is a bottom-up approach that starts with each data point as an individual cluster and merges the most similar clusters in successive iterations. By utilizing linkage criteria (e.g., Ward's method), it minimizes the increase in within-cluster variance during each merge, thereby preserving the overall structure of the data.

Spectral Clustering [15] is a graph-based clustering technique that uses the eigenvalues and eigenvectors of a similarity matrix to perform dimensionality reduction. Clustering is then performed in this reduced space, which enables the method to capture complex, non-convex cluster structures that are difficult to detect with traditional methods.

Gaussian Mixture Model (GMM) [2] is a probabilistic clustering approach that models the data as a mixture of Gaussian distributions. Using the Expectation-Maximization algorithm, GMM iteratively estimates the parameters of the Gaussian components and assigns soft membership probabilities to each data point, thus allowing for more flexible cluster boundaries.

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) [28] is a scalable clustering method designed for large datasets. It incrementally builds a clustering feature (CF) tree that summarizes the data, and subsequently applies a clustering algorithm to the CF tree, thereby reducing the computational cost while still capturing the underlying cluster structure.

OPTICS (Ordering Points to Identify the Clustering Structure) [1] is a density-based clustering algorithm that creates an ordering of the data points based on their reachability distances. This ordering reveals the intrinsic clustering structure without the need to specify a fixed density threshold, making OPTICS particularly effective for datasets with clusters of varying density.

2.4 Mixing Strategy

To strengthen the justification for employing Mixup for universum sample generation, we explore alternative mixing strategies such as CutMix, Noise Injection, Spherical Linear Interpolation (Slerp), and SnapMix, and evaluate their impact on performance.

Mixup [27] is a method that uses linear interpolation between two internal representations without relying on label information. The mixing coefficient comes from a Beta distribution. This process generates new samples that maintain semantic continuity and encourage smooth transitions in the latent space. Mixup enhances model adaptability to data distributions.

CutMix [26] is a strategy that employs random binary masks to combine regions from different internal representations. It selects a rectangular region from one representation and replaces it with a corresponding region from another representation. The mixed representation uses the proportional area of the replaced region. CutMix improves local feature fusion and encourages models to attend to diverse local information in unsupervised settings.

Spherical Linear Interpolation (Slerp) [16] is a technique that performs interpolation along the unit sphere. It maintains angular relationships between internal representations and supports natural transitions in high-dimensional latent spaces. Slerp benefits generative models and other unsupervised learning tasks.

SnapMix [7] is a strategy that uses unsupervised methods to locate discriminative regions within internal representations. It identifies important regions and mixes corresponding parts from different samples. SnapMix preserves semantic relevance and strengthens the capture of fine-grained features.

SaliencyMix [8] is a strategy that uses saliency information to guide the mixing of internal representations. It identifies key regions through saliency maps and fuses corresponding parts from different samples. SaliencyMix preserves semantic relevance and enhances focus on important features in unsupervised settings.

3 Rationale for Pseudo-labeling via K-means

To validate the reasonableness of using K-means [12] for pseudo-label generation, we compare its performance with that of several widely used clustering methods. As shown in Table 1, K-means achieves superior clustering performance on the CUB, WIKI, and NUS-WIDE datasets relative to alternative algorithms such as Agglomerative Clustering [19], Spectral Clustering [15], Gaussian Mixture Model (GMM) [2], BIRCH [28], and OPTICS [1]. These empirical and experimental results demonstrate that K-means produces high-quality pseudo-labels that effectively guide the subsequent learning process toward more discriminative feature representations.

4 Scalability to Multi-view Settings

To further demonstrate the scalability of our method beyond two-view scenarios, we conduct additional experiments on the Scene-15 dataset with three distinct views (PHOG, LBP, GIST). The results

Table 1: Final clustering performance using different pseudo-label clustering methods across three datasets (CUB, WIKI, and NUS-WIDE). The best results for each metric are highlighted in bold.

Clustering methods	CUB				WIKI				NUS-WIDE			
	ACC	NMI	ARI	Average	ACC	NMI	ARI	Average	ACC	NMI	ARI	Average
K-means	**85.33**	82.28	**74.17**	**80.59**	**62.12**	**55.74**	**46.35**	**54.74**	**64.81**	53.91	**42.16**	**53.63**
Agglomerative Clustering	84.03	82.60	73.25	79.96	58.21	53.92	44.67	52.27	60.92	53.48	35.15	49.85
Spectral Clustering	82.60	**82.65**	72.76	79.34	60.69	55.13	45.72	53.85	58.38	**54.05**	38.77	50.40
GMM	83.80	81.64	72.24	79.23	59.76	54.95	44.86	53.19	51.48	47.81	32.21	43.83
BIRCH	84.03	82.60	73.25	79.96	58.51	53.78	45.27	52.52	58.69	53.20	35.47	49.12
OPTICS	83.13	82.35	72.35	79.28	60.23	54.24	45.03	53.17	62.49	52.95	40.94	52.13

in Table 2 show that PAUSE consistently achieves the best performance across all three metrics, validating its effectiveness and robustness in more complex multi-view setups.

Table 2: Clustering performance on the Scene-15 dataset with three views (PHOG, LBP, GIST). The best results for each metric are highlighted in bold.

Method	ACC	NMI	ARI
MFLVC	40.15	41.41	24.11
CVCL	44.59	42.17	24.11
DCP	41.81	**45.23**	25.84
GCFAgg	44.14	43.40	23.99
DealMVC	40.02	42.99	24.16
HCN	45.20	44.52	28.18
PAUSE	**46.47**	45.01	**28.73**

5 Universum Boundary Analysis

To further validate that the synthesized universum samples lie near decision boundaries, we conduct a quantitative analysis based solely on distance metrics. Specifically, we measure three types of average distances: (1) from real samples to their own cluster center, (2) from universum samples to their nearest cluster center, and (3) from each cluster center to the nearest sample belonging to a different class. As summarized in Table 3, universum samples are farther from their associated cluster centers than real samples, yet significantly closer than samples from other classes. These results support that universum samples are indeed located near decision boundaries.

6 Additional Parameter Analysis

To offer additional parameter analysis beyond what is presented in the manuscript, we highlight the sensitivity of our method to each parameter and validate the selected settings through consistent performance improvements. The following sections present detailed experimental results and discussions.

6.1 Additional Analysis on λ Settings

To further support the λ analysis presented in the main text, we provide additional data in this supplementary material. Table 4 reports the clustering performance on three datasets (WIKI, NUS-WIDE, and CUB) for different constant λ settings. These detailed

Table 3: Distance-based analysis to validate the placement of universum samples.

Metric	CUB	WIKI
Real sample → own cluster center (avg.)	8.9	5.7
Universum → nearest cluster center (avg.)	12.9	7.0
Cluster center → nearest other-class sample (avg.)	36.9	18.8

Metric	NUS-WIDE
Real sample → own cluster center (avg.)	3.5
Universum → nearest cluster center (avg.)	4.5
Cluster center → nearest other-class sample (avg.)	16.5

Metric	MNIST-USPS
Real sample → own cluster center (avg.)	12.6
Universum → nearest cluster center (avg.)	17.8
Cluster center → nearest other-class sample (avg.)	87.0

results offer robust empirical validation for our parameter tuning strategy. In particular, the numerical evidence clearly demonstrates that setting $\lambda = 0.5$ yields the best performance across all metrics, thereby reinforcing its selection over other values.

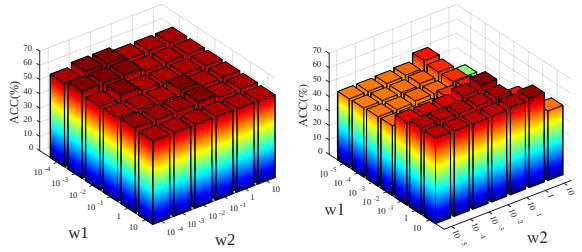
Table 4: Clustering performance on the three datasets for different Mixup settings, with λ set to constant values. The best results for each metric are highlighted in bold.

λ	WIKI			NUS-WIDE			CUB		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
0.1	55.89	54.91	41.19	59.91	52.93	36.04	83.33	81.28	73.17
0.3	61.58	55.44	45.45	62.86	53.17	38.68	85.02	82.21	73.41
0.5	**62.12**	**55.74**	**46.35**	**64.81**	**53.91**	**42.16**	**85.23**	**82.32**	**73.91**
0.7	62.04	55.49	45.68	64.29	53.84	40.70	85.10	82.13	73.69
0.9	61.81	55.46	45.38	63.78	53.59	40.02	84.93	81.97	73.22

6.2 Warm-up Parameter Analysis: w_1 and w_2

To validate the effectiveness of our warm-up parameter selection, we conduct a sensitivity analysis on the contrastive loss weights

w_1 and w_2. We vary w_1 and w_2 over $\{10^{-5}, 10^{-4}, \ldots, 10\}$ and observe the clustering performance on WIKI and NUS-WIDE datasets (Fig. 1). The WIKI dataset shows stable performance across a wide range of settings, while NUS-WIDE exhibits fluctuations when $w_1 < 0.1$ or $w_2 > 0.5$. Based on these observations, we select $w_1 = 0.3$ and $w_2 = 0.2$ for optimal performance. These findings confirm the robustness of our approach and validate the chosen parameter values for warm-up stages.

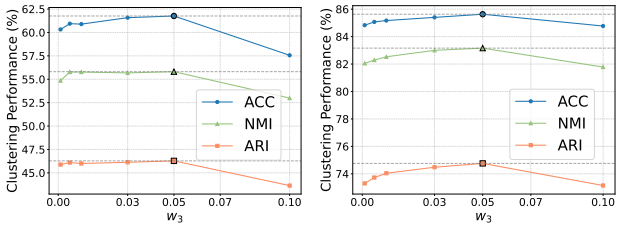


(a) WIKI (b) NUS-WIDE

Figure 1: Sensitivity analysis of the parameters w_1 and w_2 on clustering performance (ACC) for the WIKI and NUS-WIDE datasets.

6.3 Fine-tuning Parameter Analysis: w_3 and w_4

To further validate our parameter selection in the fine-tuning stage, we present an in-depth analysis of the hyperparameters w_3 and w_4. While the main text provides a 3D bar plot highlighting a region of low sensitivity, Fig. 2 and 3 offer a more fine-grained examination within this region. Our results indicate that setting $w_3 = 0.05$ and $w_4 = 0.02$ consistently yields superior clustering performance on both the WIKI and CUB datasets. This empirical evidence substantiates our chosen parameter configuration and confirms its effectiveness in enhancing the model's performance during the fine-tuning stage.

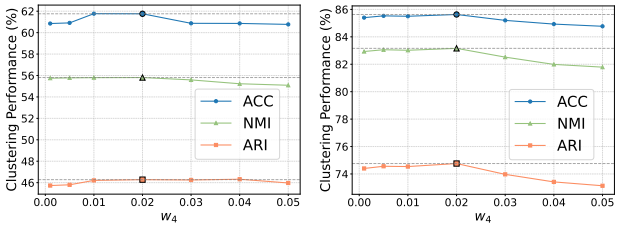


(a) WIKI (b) CUB

Figure 2: Performance analysis of the intra-view UniLoss weight parameter w_3 for the fine-tuning stage on the WIKI and CUB datasets.

7 Ablation Study on the Warm-up Stage

To provide a more complete understanding of the warm-up stage, we conduct additional ablation analysis not fully detailed in the main text. Table 5 reports the clustering performance on three datasets (CUB, WIKI, and NUS-WIDE) under various loss settings.



(a) WIKI (b) CUB

Figure 3: Performance analysis of the intra-view UniLoss weight parameter w_4 for the fine-tuning stage on the WIKI and CUB datasets.

The results clearly indicate that the joint incorporation of the reconstruction loss ($\mathcal{L}_{\text{re}}$), intra-view contrastive loss ($\mathcal{L}_{\text{intra}}$), and inter-view contrastive loss ($\mathcal{L}_{\text{inter}}$) yields the best overall performance. This empirical evidence further supports our design choice and highlights the effectiveness and robustness of our warm-up stage.

8 Robustness Analysis

To verify the robustness of our model, we compare the training stability and clustering performance of different approaches: three existing methods (DIVIDE, HCN, and ProImp) and two variants of our PAUSE framework in the second training stage—one utilizing the traditional InfoNCE loss, and the other incorporating our proposed UniLoss. As shown in Fig. 4, PAUSE begins with a warm-up stage to extract preliminary features, laying a solid foundation for subsequent optimization. However, in the fine-tuning stage, the standard InfoNCE loss is more susceptible to false negatives, often resulting in performance degradation and training instability. In contrast, the UniLoss variant maintains consistently higher and more stable accuracy across the training epochs. Moreover, while the competing baselines demonstrate increasing volatility over time, PAUSE with UniLoss shows a steady convergence toward superior clustering performance. These empirical results confirm that our method effectively alleviates the adverse effects of false negatives and overfitting, thus enhancing the robustness and reliability of the overall framework.

9 Additional Visualization Results

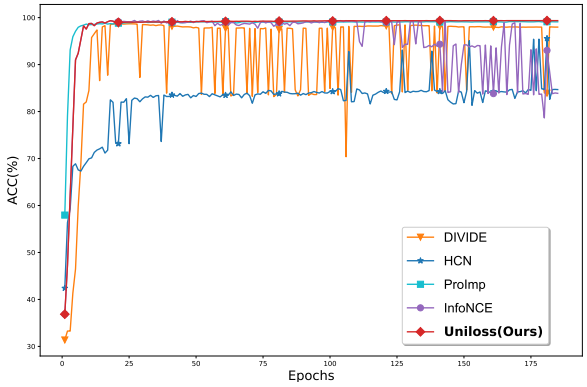
To further illustrate the clustering structures learned by our method, we provide additional t-SNE visualizations on the MNIST-USPS, NoisyMNIST and WIKI datasets, as shown in Fig. 6, 5 and 7, respectively. In addition, Fig. 8 presents visualization results from our ablation studies on the UniLoss mechanism using the NUS-WIDE dataset. These visualizations provide qualitative evidence supporting the effectiveness of our method in producing well-separated and semantically meaningful clusters across diverse datasets.

10 Additional Theoretical Analysis of UniLoss

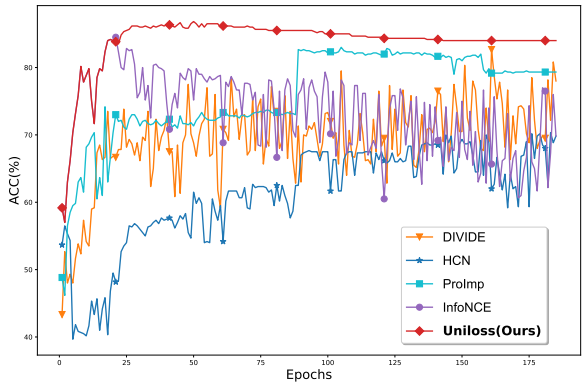
To provide deeper theoretical insights into the design and effectiveness of our proposed UniLoss, this section presents additional analytical discussions and gradient-based evaluations.

Table 5: Clustering performance in the warm-up stage under different loss settings on three datasets (CUB, WIKI, and NUS-WIDE). A check mark (✓) indicates the inclusion of the corresponding module. Note that the contrastive losses used during the warm-up stage are traditional contrastive losses.

Stage	$\mathcal{L}_{re}$	$\mathcal{L}_{intra}$	$\mathcal{L}_{inter}$	CUB			WIKI			NUS-WIDE		
				ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Warm-up	✓			64.00	64.28	47.09	48.92	47.83	25.80	52.68	40.61	30.15
		✓		60.53	59.42	33.63	49.72	48.55	20.89	28.17	22.14	13.01
			✓	67.83	68.54	51.07	53.81	49.37	26.33	57.72	50.96	31.68
		✓	✓	72.67	70.85	51.46	54.40	50.45	33.77	58.69	49.91	33.47
	✓	✓		67.17	66.87	46.51	54.76	52.01	36.17	43.46	41.00	17.21
	✓		✓	68.17	72.75	58.03	54.43	50.01	30.13	58.36	50.27	33.59
	✓	✓	✓	**75.17**	**74.07**	**60.98**	**55.35**	**53.03**	**40.09**	**60.36**	**51.50**	**35.13**



(a) MNIST-USPS



(b) CUB

Figure 4: Clustering accuracy (ACC) curves on the MNIST-USPS and CUB datasets over 180 epochs.

10.1 Gradient Derivation

To rigorously analyze the properties of the proposed inter-view UniLoss, we first derive its gradient with respect to the anchor representation z_i^1. Recall that the boundary term $D_U(z_i^1, z_i^2, \mathcal{U}_{12})$ is defined as:

$$D_U(z_i^1, z_i^2, \mathcal{U}_{12}) = \sum_{U_j \in \mathcal{U}_{12}} \exp\left(\frac{S(z_i^1, U_j)}{\tau}\right) + \sum_{U_j \in \mathcal{U}_{12}} \exp\left(\frac{S(z_i^2, U_j)}{\tau}\right). \tag{1}$$

Since the second term does not depend on z_i^1, its derivative vanishes. Thus, the gradient with respect to z_i^1 is:

$$\nabla_{z_i^1} D_U(z_i^1, z_i^2, \mathcal{U}_{12}) = \frac{1}{\tau} \sum_{U_j \in \mathcal{U}_{12}} \exp\left(\frac{z_i^{1\top} U_j}{\tau}\right) U_j. \tag{2}$$

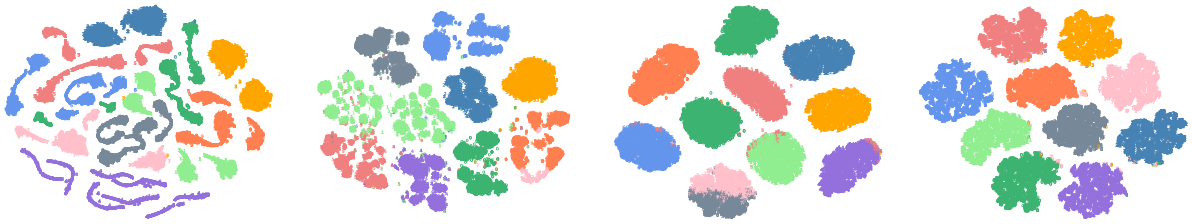
10.2 Gradient Directionality Analysis.

We next examine the geometric implications and theoretical soundness of the derived gradient. The expression clearly indicates a weighted sum over all universum samples, $U_j \in \mathcal{U}_{12}$. Specifically, each universum sample U_j contributes proportionally to its exponential similarity with the anchor z_i^1. Formally, the gradient can be decomposed as:

$$\nabla_{z_i^1} D_U(z_i^1, z_i^2, \mathcal{U}_{12}) = \frac{1}{\tau} \sum_{U_j \in \mathcal{U}_{12}} \alpha_{ij} U_j, \quad \text{where} \quad \alpha_{ij} = \exp\left(\frac{z_i^{1\top} U_j}{\tau}\right). \tag{3}$$

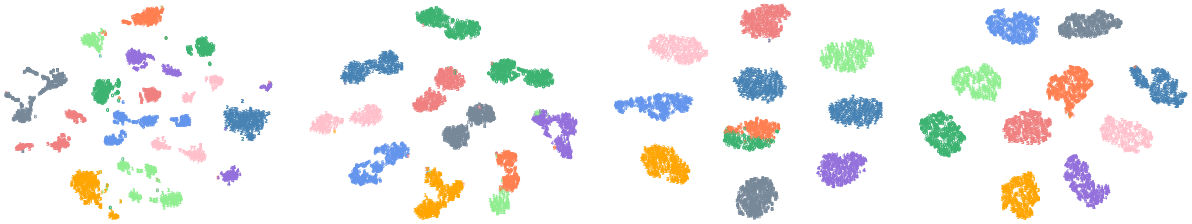
The coefficient α_{ij} explicitly quantifies how strongly each universum sample repels the anchor. Because the exponential function monotonically increases with similarity, universum samples closer to the anchor representation (larger $z_i^{1\top} U_j$) exert greater repulsive forces. Therefore, the gradient inherently embodies a dynamic hard-negative mining mechanism, primarily pushing the anchor away from universum points that are semantically ambiguous or visually close, effectively establishing a neutral inter-class region around each class cluster.

Geometrically, assuming the representations are normalized onto the unit hypersphere, we have $z_i^{1\top} U_j = \cos(\theta_{ij})$, where θ_{ij} denotes the angle between z_i^1 and U_j. Hence, universum samples with smaller angular distances contribute significantly more, ensuring that the gradient direction strongly repels the anchor from boundary-confusing areas. This design reduces false negatives by avoiding inadvertent repulsion among semantically close same-class anchors.



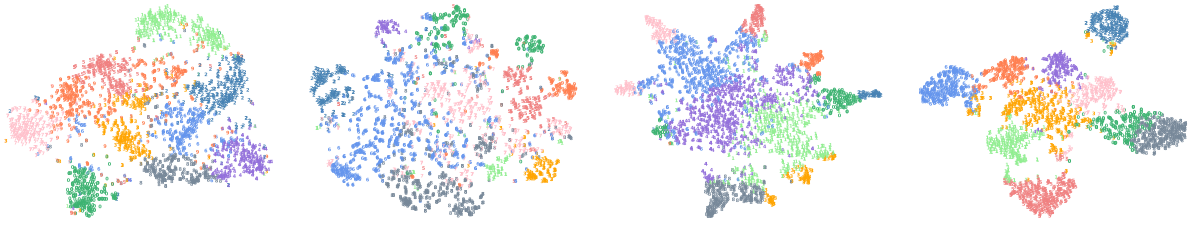
(a) DIVIDE (b) HCN (c) ProImp (d) PAUSE(Ours)

Figure 5: The t-SNE visualization results of the NoisyMNIST dataset using different methods. Data points are colored according to their respective classes.



(a) DIVIDE (b) HCN (c) ProImp (d) PAUSE(Ours)

Figure 6: The t-SNE visualization results of the MNIST-USPS dataset using different methods. Data points are colored according to their respective classes.



(a) DIVIDE (b) HCN (c) ProImp (d) PAUSE(Ours)

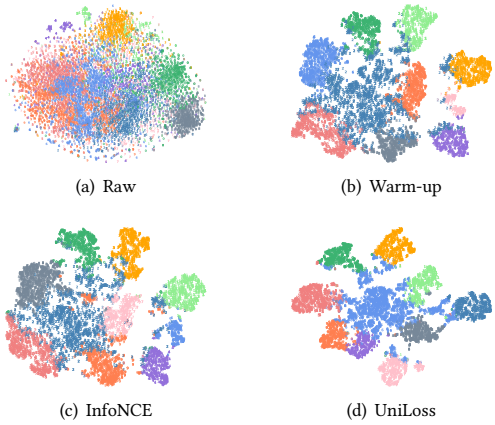
Figure 7: The t-SNE visualization results of the WIKI dataset using different methods. Data points are colored according to their respective classes.

10.3 Boundary Maximization

Explicitly incorporating universum samples in the UniLoss significantly reinforces boundary maximization. By construction, universum samples reside precisely in inter-class spaces, outlining decision boundaries clearly. Optimizing UniLoss effectively enlarges the margins between class clusters by continuously pushing class-specific anchors away from universum-defined boundary regions. This intrinsic property systematically enlarges inter-class margins, thereby enhancing model discriminability and robustness.

10.4 Convergence Analysis

Finally, we discuss the convergence characteristics of UniLoss. Structurally similar to established contrastive objectives (e.g., InfoNCE), UniLoss maintains continuity, differentiability, and possesses a lower-bounded loss landscape. Under standard optimization conditions (appropriate learning rates and sufficiently smooth similarity metrics), gradient-based optimization methods guarantee monotonic descent and stable convergence to stationary points. The inclusion of universum boundary constraints maintains these desirable convexity properties locally, ensuring stable training dynamics and convergence to well-defined, discriminative feature representations.



(a) Raw (b) Warm-up

(c) InfoNCE (d) UniLoss

Figure 8: The t-SNE visualizations on NUS-WIDE: (a) raw data; (b) after warm-up training; (c) warm-up followed by intra- and inter-view InfoNCE; (d) warm-up followed by intra- and inter-view UniLoss. Subfigures (c) and (d) directly compare the traditional InfoNCE loss with the proposed UniLoss under identical conditions.

10.5 Intra-Class Consistency Enhancement

When semantically similar samples are incorrectly repelled, the resulting latent representations fail to accurately capture the clustering assignment information, which increases the conditional entropy $H(Z|Y)$. An increase in $H(Z|Y)$ reflects a weaker dependence between the latent representations and cluster assignments, thus diminishing the model's discriminative power. By introducing universum samples positioned near the decision boundary between classes, we prevent semantically similar samples from being erroneously classified as negatives. These neutral universum samples reduce uncertainty in negative sample selection, stabilize the latent representations, and minimize intra-class overlap. As a result, the relationship between latent representations and clustering assignments strengthens, improving the model's ability to distinguish between classes. Mathematically, this effect is captured by the mutual information equation: $I(Z;Y) = H(Z) - H(Z|Y)$. By reducing $H(Z|Y)$, we increase $I(Z;Y)$, which enhances intra-class consistency and improves the robustness and discriminative ability of the model.

10.6 Inter-Class Separation Enhancement

Traditional contrastive learning methods often suffer from poor negative sample selection, leading to an increase in conditional entropy $H(Z|Y)$, which causes blurred class boundaries. Inadequate negative sample selection can lead to semantically similar samples being incorrectly treated as negatives, resulting in indistinct class boundaries and poor inter-class separation. The introduction of universum samples addresses this issue by stabilizing negative sample selection and creating clearer decision boundaries between classes. These universum samples, located at the inter-class boundary, neither belong to any specific class nor are misclassified as negative samples. As a result, the model achieves sharper class boundaries, reducing overlap between classes and improving inter-class separability. Additionally, in multi-view learning, universum samples enhance the optimization of inter-class separation in the latent space by increasing the distance between different classes. This enhanced separation improves the model's ability to generalize, particularly in noisy label scenarios, as it allows the model to more effectively differentiate between classes. The reduction in conditional entropy $H(Z|Y)$ further strengthens the relationship between latent representations and class labels, improving inter-class separability and enhancing the model's discriminative power.

References

[1] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. 1999. OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod Record* 28, 2 (1999), 49–60.
[2] Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*. Vol. 4. Springer.
[3] Jie Chen, Hua Mao, Wai Lok Woo, and Xi Peng. 2023. Deep multiview clustering by contrasting cluster assignments. In *2023 IEEE/CVF International Conference on Computer Vision*. 16706–16715.
[4] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*. 1–9.
[5] Ruiming Guo, Mouxing Yang, Yijie Lin, Xi Peng, and Peng Hu. 2024. Robust contrastive multi-view clustering against dual noisy correspondence. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*, Vol. 37.
[6] Changhao He, Hongyuan Zhu, Peng Hu, and Xi Peng. 2024. Robust variational contrastive learning for partially view-unaligned clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 4167–4176.
[7] Shaoli Huang, Xinchao Wang, and Dacheng Tao. 2021. Snapmix: semantically proportional mixing for augmenting fine-grained data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1628–1636.
[8] Taesung Kim, Minkyu Kim, Jeongsoo Park, and Nojun Kwak. 2021. Saliencymix: a saliency guided data augmentation strategy for better regularization. In *International Conference on Learning Representations*.
[9] Haobin Li, Yunfan Li, Mouxing Yang, Peng Hu, Dezhong Peng, and Xi Peng. 2023. Incomplete multi-view clustering via prototype-based imputation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. 3911–3919.
[10] Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. 2022. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2022), 4447–4461.
[11] Yiding Lu, Yijie Lin, Mouxing Yang, Dezhong Peng, Peng Hu, and Xi Peng. 2024. Decoupled contrastive multi-view clustering with high-order random walks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 14193–14201.
[12] James MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, Vol. 5. 281–298.
[13] Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. 2019. COMIC: multi-view clustering without parameter selection. In *International Conference on Machine Learning*. 5092–5101.
[14] Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2013. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 3 (2013), 521–535.
[15] Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 8 (2000), 888–905.
[16] Ken Shoemake. 1985. Animating rotation with quaternion curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*. 245–254.
[17] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. (2011).
[18] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On deep multi-view representation learning. In *International Conference on Machine Learning*. 1083–1092.
[19] Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *J. Amer. Statist. Assoc.* 58, 301 (1963), 236–244.
[20] Chengwei Xia, Chaoxi Niu, and Kun Zhan. 2025. Hierarchical consensus network for multiview feature learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 21617–21625.

[21] Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. 2022. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16051–16060.

[22] Weiqing Yan, Yuanyang Zhang, Chenlei Lv, Chang Tang, Guanghui Yue, Liang Liao, and Weisi Lin. 2023. Gcfagg: global and cross-view feature aggregation for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19863–19872.

[23] Mouxing Yang, Yunfan Li, Peng Hu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. 2022. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2022), 1055–1069.

[24] Mouxing Yang, Yunfan Li, Zhenyu Huang, Zitao Liu, Peng Hu, and Xi Peng. 2021. Partially view-aligned representation learning with noise-robust contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1134–1143.

[25] Xihong Yang, Jin Jiaqi, Siwei Wang, Ke Liang, Yue Liu, Yi Wen, Suyuan Liu, Sihang Zhou, Xinwang Liu, and En Zhu. 2023. Dealmvc: dual contrastive calibration for multi-view clustering. In *Proceedings of the 31st ACM International Conference on Multimedia*. 337–346.

[26] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6023–6032.

[27] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. Mixup: beyond empirical risk minimization. In *International Conference on Learning Representations*.

[28] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996. BIRCH: an efficient data clustering method for very large databases. *ACM Sigmod Record* 25, 2 (1996), 103–114.

[29] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. 2019. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10394–10403.