SeqMvRL: A Sequential Fusion Framework for Multi-view Representation Learning

Supplementary Material

A. Reward

The proposed reward is designed to motivate the agent to select views that enhance clustering performance by assessing both intra-cluster and inter-cluster distances, along with assignment accuracy. The reward is calculated as follows:

$$r = (1-\gamma)\left(\frac{\alpha}{1+d_{\text{intra}}} + \beta \cdot d_{\text{inter}}\right) + \gamma \cdot (2 \cdot \delta_{\hat{y}_i, y_i} - 1), \tag{A.1}$$

where d_{intra} measures the intra-cluster distance between the view and its assigned cluster center, favoring smaller values, and d_{inter} denotes the inter-cluster distance between the view and other cluster centers, with larger values being desirable. The parameters α and β balance the contributions of these two distance metrics, while $\gamma \in [0, 1]$ determines the relative influence of clustering quality and assignment accuracy. The indicator function $\delta_{\hat{y}_i, y_i}$ takes a value of 1 if the predicted cluster label $\hat{y}_i$ matches the true label y_i, and 0 otherwise. A higher reward value indicates that the current view is closer to its assigned cluster center and farther from other cluster centers, providing greater benefit for clustering optimization. Figure 6 illustrates an example of this reward.

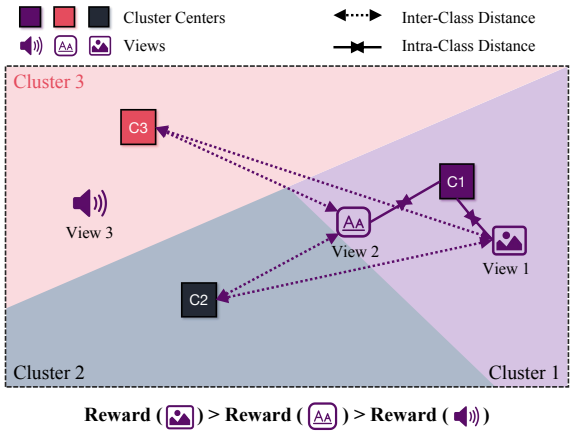

Figure 6. Illustration of the proposed reward. Take a sample from cluster 1 with three views as an example, and give three cluster centroids: View 1 has a smaller intra-cluster distance and a larger inter-cluster distance than View 2. View 3 is assigned to the incorrect cluster. As a result, the calculated reward values follow the order: View 1 > View 2 > View 3.

A.1. Balance Parameter γ

As a supplement to Figure 4 of the main manuscript, we explored the impact of γ on the reward function. Figure 7 demonstrates how the choice of γ affects the balance between clustering quality and assignment accuracy in the reward function. As γ increases, the emphasis shifts from optimizing clustering distances to prioritizing correct cluster assignments. This trade-off is evident in the performance metrics. A small γ value amplifies the error caused by incorrect assignments, while a large γ leads to a lack of fine-grained distance supervision, slightly reducing accuracy.

A.2. Cluster Center

In the proposed reward strategy, the accuracy of cluster center is critical. To mitigate the impact of outliers on the cluster center, we compute it using a certain proportion of the training data. Figure 8 shows the results of calculating cluster centers using different proportions of the data. The figure highlights the trade-off between robustness and representativeness as the proportion of data changes. Using a smaller proportion reduces the influence of outliers, resulting in more stable cluster centers, but it may lead to fewer representative centers for the entire dataset. Conversely, using a larger proportion increases representativeness but risks incorporating more noise from outliers. We select 90% of the data to balance stability and representativeness.

A.3. Maximum Sequence Length T

Maximum sequence length T in Algorithm 1 is a common hyper-parameter in reinforcement learning, balancing computational cost and performance. To study its impact, we conducted experiments without a predefined maximum length, relying solely on the 'END' action to stop fusion. The final sequence length and its frequency are shown in Figure 10. Longer sequences occur less frequently but incur higher computational costs. Thereby, a maximum length T is an efficient strategy to avoid unnecessary costs. Figure 9 further shows the accuracy for different sequence lengths on the COIL-20 dataset with three views. A clear trend emerges: overly long sequences lead to performance degradation, likely due to an increased risk of introducing conflicting information. Therefore, we set the maximum sequence length to 1.5 times the number of views.

A.4. Learned Sequences

Besides the results in Figure 3 of the main manuscript, we further analyze view sequences by comparing our learned sequence with 4 random ones across fusion steps, in terms

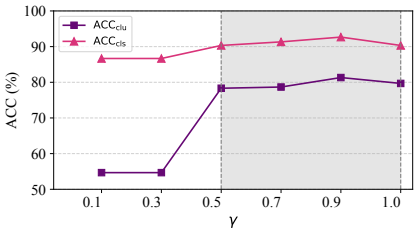

Figure 7. Accuracy trend for different γ values on the COIL-20 dataset.

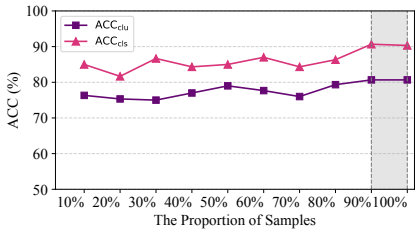

Figure 8. Accuracy trend for different data proportions on the COIL-20 dataset.

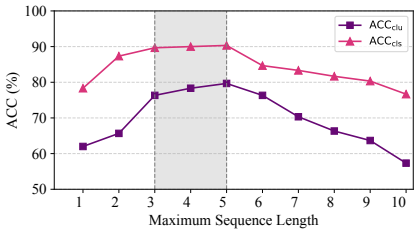

Figure 9. Accuracy trend for different maximum lengths T on the COIL-20 dataset.

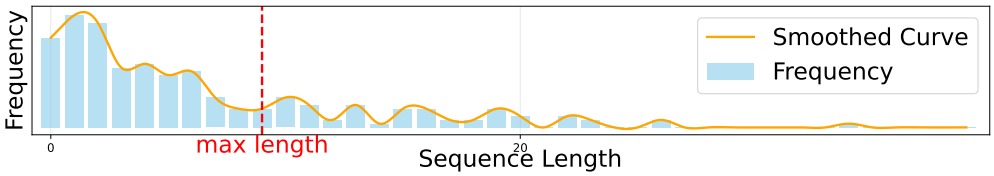

Figure 10. Comparison of the frequency of occurrence of the final sequence length in the absence of a predefined maximum length.

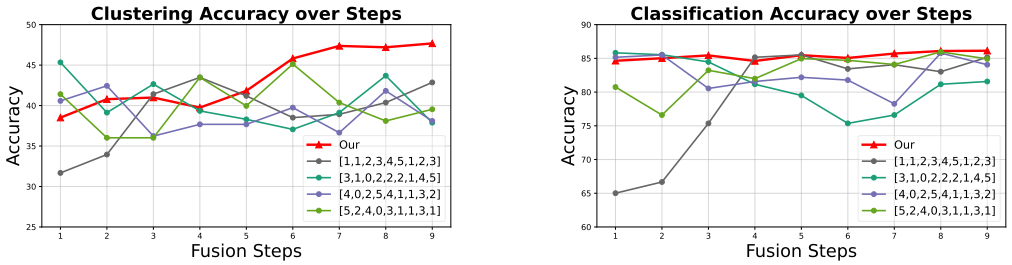

Figure 11. Comparison of learned and random sequences across fusion steps in terms of clustering (left) and classification (right) accuracy.

of clustering and classification accuracy, shown in Figure 11. Random sequences fluctuate significantly due to view conflicts, while our learned sequence progressively improves and eventually stabilizes, indicating effective view selection for improved fused representation quality.

A.5. Sequence Initialization Strategies

We compare four initialization strategies on Caltech-20, with clustering accuracy shown in Table 6. We select random initialization as it balances performance and cost while enhancing NVS generalization and stability.

Default view	Max-dimension view	Best view	**Random view**
36.65	46.79	48.88	47.85
dataset-dependent, unstable	requires manual selection	requires extra model training	balances performance and cost

Table 6. Comparison of Different Initialization Strategies.

A.6. Reward functions

We compare four reward functions: cluster accuracy δ, inter-cluster distance d_{inter}, intra-cluster distance d_{intra}, and our proposed reward function r, as shown in Table 7. The results show that r achieves the highest accuracy on Caltech-20, demonstrating its effectiveness.

Caltech-20	δ	d_{inter}	d_{intra}	r **(our)**
ACC_{clu}	46.79	19.46	26.29	**47.85**
ACC_{cls}	85.51	61.28	62.11	**86.13**

Table 7. Comparison of different reward functions

B. Results in Various Representative Scenarios

To validate the performance of the proposed parallel fusion framework, we designed three challenging scenarios: numerous views, low-quality views, multi-modal data. These scenarios emphasize the robustness and flexibility of sequential fusion across diverse and complex conditions.

B.1. Numerous Views

The COIL20-v20 Dataset with 20 Views. Building upon the COIL-20 dataset, we constructed a new dataset, COIL20-v20, comprising 20 views to evaluate SeqMvRL's

View-ID	Feature	View Dimensions	Describe
View-1	HOG	320	Captures edges and gradients by analyzing gradient directions.
View-2	LBP	512	Encodes texture based on pixel intensity patterns.
View-3	Zernike	25	Extracts shape features with rotational invariance.
View-4	Gabor	1024	Detects edges and textures at different scales and angles.
View-5	Haralick	13	Computes texture statistics like contrast and homogeneity.
View-6	Fourier	1024	Analyzes frequency patterns for global image features.
View-7	Wavelet	1024	Breaks the image into different scales and frequencies.
View-8	Gray	32	Uses intensity values from grayscale images.
View-9	SIFT	512	Detects and describes local keypoints with scale and rotation invariance.
View-10	Harris	1024	Identifies corner points based on local intensity changes.
View-11	ResNet-18	512	A simple 18-layer network with skip connections to avoid vanishing gradients.
View-12	ResNet-34	512	A deeper 34-layer version of ResNet for better feature learning.
View-13	ResNet-50	2048	A 50-layer network using bottleneck blocks.
View-14	VGG-16	4608	A straightforward 16-layer model with a uniform structure.
View-15	VGG-19	4608	A 19-layer version of VGG with slightly better accuracy.
View-16	DenseNet-121	1024	A 121-layer model where all layers connect for better gradient flow.
View-17	MobileNet-v2	1280	A lightweight model optimized for mobile devices with efficient convolutions.
View-18	EfficientNet	1280	A model family scaling depth, width, and resolution.
View-19	Inception-v3	2048	A network with inception modules to capture multi-scale features efficiently.
View-20	AlexNet	1024	An early 8-layer network introducing ReLU and dropout.

Table 8. Construction Method, Dimensionality, and Description of Each View in the 20-View Dataset.

	ACC_{clu}	NMI	ARI	ACC_{cls}	Prec	F-score
DSMVC	79.33	88.53	76.35	94.67	95.67	94.72
SeqMvRL (**Our**)	**86.75**	**91.52**	**81.33**	**97.67**	**98.01**	**97.64**
Δ SOTA	↑ 7.42	↑ 2.99	↑ 4.98	↑ 3.00	↑ 2.34	↑ 2.92

Table 9. Results on COIL20-v20 datasets.

performance in numerous views scenarios. The first 10 views are generated using traditional feature extraction methods, capturing classical characteristics of the data. The remaining 10 views are derived from various deep networks, providing richer and more abstract representations. Table 8 summarizes the construction method, dimensionality, and a brief description of each view.

Results on the COIL20-v20 Dataset. Table 9 presents the results on the COIL20-v20 dataset. The results indicate that, compared to datasets with a smaller number of views, the proposed SeqMvRL achieves significant improvements in both clustering and classification performance (e.g., **7.42** on ACC_{clu}) in scenarios with a large number of redundant views. A likely reason for this is that sequential fusion effectively reduces the impact of redundancy while emphasizing the most informative features. These findings highlight the effectiveness of sequential fusion in handling datasets with high view overlap, ensuring more robust and reliable clustering and classification outcomes.

Different Numbers of Views. To further explore the impact of the number of views on clustering performance, we conducted experiments as the number of views varies from 5 to 20. The results in Figure 12 show that as the number of views increases, the clustering performance initially improves due to the integration of complementary information from additional views. However, after 13 views, the performance gain begins to plateau, and slight degradation is observed in some cases beyond 16 views, likely due to the increasing redundancy and noise introduced by additional views. The TSNE plots shown in Figure 13 also validate this observation. These findings highlight the robustness of our method in leveraging complementary information while maintaining resilience to redundancy.

B.2. Low-Quality Views

In a similar manner, based on COIL20, we further constructed a COIL20-Noise dataset with low-quality views. Specifically, we treated the image captured every 30 degrees in COIL20 as a new sample. Each sample consists of six views, and we further added random noise of varying intensities to simulate low-quality views in real-world scenarios. Table 10 compares the clustering accuracy ACC_{clu} of four additional state-of-the-art methods with the first two under incremental learning. $CAC_{\mathbf{SeqMvRL}}$ enhances CAC incremental fusion using our learned view orders. SeqMvRL surpasses all baselines on most datasets, significantly enhancing incremental fusion with its learned orders. The table also reports training/inference time per sample on COIL-20, showing that SeqMvRL achieves significant performance gains with an acceptable increase in time overhead.

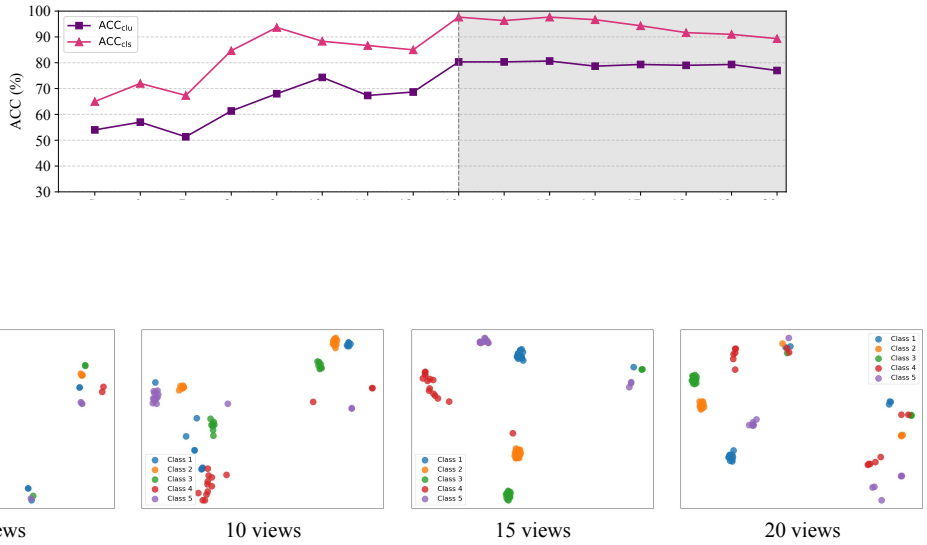

Figure 13. T-SNE visualization using 5, 10, 15, and 20 views.

Table 10. Comparison with additional state-of-the-art methods.

	Time per sample	COIL-20	COIL20-Noise
LAIMVC$_{MM24}$	- / 1.40ms	72.71	60.13
CAC$_{AAAI24}$	2.19s / 1.59ms	77.02	67.29
CAC$_{\textbf{SeqMvRL}}$	2.22s / 1.60ms	80.54	71.43
MRDD$_{CVPR24}$	1.2s / 3.77ms	72.19	64.57
ESTMC$_{TPAMI25}$	- / 41.68ms	77.36	66.88
SeqMvRL	0.90s / 1.03ms	79.68	66.95

B.3. Details of Multi-Modal Experiments

Multi-modal and multi-view data both provide multiple representations of the same samples but differ in their sources and characteristics. Multi-view data arises from a single modality but different perspectives or techniques, such as varying imaging angles or feature extraction methods. In contrast, multi-modal data integrates fundamentally different sources like text, images, or audio, requiring methods to bridge semantic and structural gaps between modalities, making it more challenging. The details of the multi-modal data used in Table 4 are provided below.

- **Oxford-IIIT Pet** [52] is a widely used benchmark dataset in computer vision, featuring 37 distinct categories that include 25 cat breeds and 12 dog breeds. The dataset contains a total of 7,390 images, with approximately 200 images per category. Each image is annotated with class labels and pixel-level segmentation masks, enabling tasks such as classification and segmentation. The dataset provides a variety of poses, lighting conditions, and occlusions, making it suitable for evaluating the robustness of models in real-world scenarios.
- **CAT** [53] is designed for keypoint detection and pose estimation, containing 9,997 images of cats across 7 categories. Each image is annotated with 9 key points (left and right eyes, mouth, and 3 points for each ear), represented as (x, y) pixel coordinates, resulting in 18 dimensions per annotation. The dataset includes variations in pose, lighting, and background, making it suitable for evaluating model robustness in keypoint detection and classification.

For image data, we used ResNet50 pre-trained on ImageNet to extract features. The fully connected layer was removed, and the output from the global average pooling (GAP) layer served as the initial feature representation. Images were resized to 224×224 pixels and normalized based on ImageNet's mean and standard deviation, resulting in a 2048-dimensional feature vector for each image.

For text data, we used DistilBERT, a lightweight transformer-based model. The text was tokenized, padded, and truncated to a maximum length of 128 tokens. The output from DistilBERT's last layer was averaged to produce a 768-dimensional feature vector for each text description.

For keypoint data, they were flattened into an 18-dimensional vector, providing a compact geometric representation of each sample.