Multi-Hierarchical Contrastive Spectral Fusion for Multi-View Clustering

Supplementary Material

A. Theoretical Analysis

A1. Proof for Theorem 1

Theorem 1. *Minimizing the spectral loss $\mathcal{L}_{spc}$ is equivalent to minimizing standard spectral embedding objective $\sum_v Tr(\mathbf{P}^{(v)\top}\mathbf{L}^{(v)}\mathbf{P}^{(v)})$.*

Proof. For each view v, the spectral loss is defined as:

$$\mathcal{L}_{spc} = \sum_{i,j} W_{ij}^{(v)} \|\mathbf{p}_i^{(v)} - \mathbf{p}_j^{(v)}\|^2, \tag{16}$$

where $W_{ij}^{(v)}$ denotes the similarity between samples i and j in view v, and $\mathbf{p}_i^{(v)}$ is the embedding of sample i in view v.

Let $\mathbf{L}^{(v)} = \mathbf{D}^{(v)} - \mathbf{W}^{(v)}$ be the graph Laplacian for view v, where $\mathbf{D}^{(v)}$ is the degree matrix. We have:

$$\begin{aligned}
&\sum_{i,j} W_{ij}^{(v)} \|\mathbf{p}_i^{(v)} - \mathbf{p}_j^{(v)}\|^2 \\
&= \sum_{i,j} W_{ij}^{(v)} (\|\mathbf{p}_i^{(v)}\|^2 + \|\mathbf{p}_j^{(v)}\|^2 - 2\mathbf{p}_i^{(v)\top}\mathbf{p}_j^{(v)}) \\
&= 2\sum_i D_{ii}^{(v)} \|\mathbf{p}_i^{(v)}\|^2 - 2\sum_{i,j} W_{ij}^{(v)} \mathbf{p}_i^{(v)\top}\mathbf{p}_j^{(v)} \\
&= 2 \cdot \mathrm{Tr}(\mathbf{P}^{(v)\top}\mathbf{L}^{(v)}\mathbf{P}^{(v)}).
\end{aligned} \tag{17}$$

Thus, minimizing $\mathcal{L}_{spc}$ is equivalent to minimizing $\sum_v \mathrm{Tr}(\mathbf{P}^{(v)\top}\mathbf{L}^{(v)}\mathbf{P}^{(v)})$, which is the standard objective of spectral embedding. This ensures that samples connected with high similarity ($W_{ij}^{(v)}$) remain close in the embedding space $\mathbf{P}^{(v)}$, preserving the local geometric structure. □

A2. Proof that MCSF Satisfies Definition 1

Definition 1. ***(Deep Spectral Embedding)*** *Given a graph Laplacian $\mathbf{L} \in \mathbb{R}^{n\times n}$, A representation $\mathbf{P}^{(v)} \in \mathbb{R}^{n\times K}$ learned via a neural network is a deep spectral embedding if it satisfies:*

1. Clustering structure preservation: $\mathbf{P}^{(v)}$ minimizes the Laplacian objective: $\mathrm{Tr}({\mathbf{P}^{(v)}}^\top \mathbf{L}^{(v)}\mathbf{P}^{(v)})$;

2. Orthonormality: ${\mathbf{P}^{(v)}}^\top \mathbf{P}^{(v)} = \mathbf{I}$;

3. End-to-end trainability: $\mathbf{P}^{(v)}$ is differentiable with respect to the network parameters.

Proof. Without loss of generality, we consider an arbitrary view for the proof.

(1) Structure Preservation. The spectral loss in MCSF is defined as:

$$\mathcal{L}_{\text{spc}} = \sum_{i,j} W_{ij} \|\mathbf{p}_i - \mathbf{p}_j\|^2. \tag{18}$$

According to Theorem 1, $\mathcal{L}_{\text{spc}}$ is equivalent (up to a scalar) to the classical spectral embedding objective $\mathrm{Tr}(\mathbf{P}^\top\mathbf{L}\mathbf{P})$.

(2) Orthonormality Constraint. Let $\mathbf{H} \in \mathbb{R}^{n\times K}$ denote the probability matrix. To ensure numerical stability, we compute the matrix:

$$\mathbf{R} = \mathrm{chol}(\mathbf{H}^\top\mathbf{H} + \epsilon\mathbf{I}), \tag{19}$$

where $\epsilon > 0$ is a small constant (*e.g.*, 10^{-5}) and $\mathrm{chol}(\cdot)$ denotes the Cholesky decomposition, which returns a lower triangular matrix $\mathbf{R}$ such that:

$$\mathbf{H}^\top\mathbf{H} + \epsilon\mathbf{I} = \mathbf{R}\mathbf{R}^\top. \tag{20}$$

The final embedding is then obtained as:

$$\mathbf{P} = \mathbf{H}(\mathbf{R}^{-1})^\top. \tag{21}$$

We now analyze the orthogonality of $\mathbf{P}$:

$$\begin{aligned}
\mathbf{P}^\top\mathbf{P} &= (\mathbf{R}^{-1})\mathbf{H}^\top\mathbf{H}(\mathbf{R}^{-1})^\top \\
&= (\mathbf{R}^{-1})(\mathbf{R}\mathbf{R}^\top - \epsilon\mathbf{I})(\mathbf{R}^{-1})^\top \\
&= (\mathbf{R}^{-1})\mathbf{R}\mathbf{R}^\top(\mathbf{R}^{-1})^\top - \epsilon(\mathbf{R}^{-1})(\mathbf{R}^{-1})^\top \\
&= \mathbf{I} - \epsilon(\mathbf{R}^{-1})(\mathbf{R}^{-1})^\top.
\end{aligned} \tag{22}$$

Let $\mathbf{M} = \mathbf{R}^{-1}$. Since $\mathbf{R}$ is a lower triangular matrix, $\mathbf{M}$ is also lower triangular, and $\mathbf{M}\mathbf{M}^\top$ is a symmetric positive definite matrix. Therefore,

$$\mathbf{P}^\top\mathbf{P} = \mathbf{I} - \epsilon\mathbf{M}\mathbf{M}^\top \approx \mathbf{I}. \tag{23}$$

Interpretation and Practical Considerations:

- The embedding $\mathbf{P}$ is not strictly orthogonal but approximately orthogonal. The deviation from orthogonality is controlled by the term $\epsilon\mathbf{M}\mathbf{M}^\top$.
- In practice, ϵ is chosen to be very small (*e.g.*, 10^{-5} or 10^{-8}) to ensure numerical stability during Cholesky decomposition without significantly violating the orthogonality constraint.
- This approximation is widely adopted in deep learning frameworks [5, 6, 10] to enforce differentiable constraints and has proven effective in practice.

Thus, for practical purposes, $\mathbf{P}$ satisfies the orthonormality constraint sufficiently well.

(3) End-to-End Learnability. The structure-preserving spectral loss for each view v in MCSF is defined as:

$$\mathcal{L}_{\text{spc}} = \sum_{i,j=1}^{n} W_{ij} \|\mathbf{p}_i - \mathbf{p}_j\|^2 = 2\,\mathrm{Tr}\left(\mathbf{P}^\top\mathbf{L}\mathbf{P}\right). \tag{24}$$

Gradient Derivation According to matrix calculus, we have:

$$\frac{\partial}{\partial \mathbf{P}} 2\,\mathrm{Tr}\left(\mathbf{P}^\top \mathbf{L}\mathbf{P}\right) = 4\mathbf{L}\mathbf{P}. \tag{25}$$

Therefore, the gradient of the spectral loss with respect to $\mathbf{P}$ is:

$$\frac{\partial \mathcal{L}_{\text{spc}}}{\partial \mathbf{P}} = 4\mathbf{L}\mathbf{P}. \tag{26}$$

Differentiability and Backpropagation Modern automatic differentiation frameworks (*e.g.*, PyTorch) support backpropagation through Cholesky decomposition and related matrix operations [9]. Since $\mathbf{P}$ is constructed via differentiable operations (matrix multiplication, inversion, and Cholesky decomposition), and $\mathbf{W}$ is computed from encoder features using differentiable kernels (*e.g.*, Gaussian or cosine), the entire spectral loss $\mathcal{L}_{\text{spc}}$ remains differentiable with respect to the encoder parameters. Thus, gradients $\nabla \mathcal{L}_{\text{spc}}$ can be efficiently computed and propagated during training.

□

A3. Proof for Theorem 2

Theorem 2. *Minimizing the multi-hierarchical contrastive loss $\mathcal{L}_c$ simultaneously maximizes the following:*

$$I(\mathbf{P}^{(v)}; \mathbf{P}^{(v)}) + I(\mathbf{P}^{(v)}; \mathbf{P}) + I(\mathbf{P}; \mathbf{P}), \tag{27}$$

where $I(\mathbf{P}^{(v)}; \mathbf{P}^{(v)})$, $I(\mathbf{P}^{(v)}; \mathbf{P})$, and $I(\mathbf{P}; \mathbf{P})$ correspond to the mutual information for VSP, VCA, and CSR, respectively.

Proof. By applying Lemma 1 to each component of $\mathcal{L}_c$ in Eq. (10), we have: For VSP: $\mathcal{L}_{VSP} \geq C_1 - I(\mathbf{P}^{(v)}; \mathbf{P}^{(v)})$; For VCA: $\mathcal{L}_{VCA} \geq C_2 - I(\mathbf{P}^{(v)}; \mathbf{P})$; For CSR: $\mathcal{L}_{CSR} \geq C_3 - I(\mathbf{P}; \mathbf{P})$. Combining these inequalities, we obtain:

$$\begin{aligned} \mathcal{L}_c \geq & (C_1 + C_2 + C_3) \\ & - \left(I(\mathbf{P}^{(v)}; \mathbf{P}^{(v)}) + I(\mathbf{P}^{(v)}; \mathbf{P}) + I(\mathbf{P}; \mathbf{P}) \right). \end{aligned} \tag{28}$$

Thus, minimizing $\mathcal{L}_c$ maximizes the sum of the mutual information terms, proving the theorem. □

A4. Proof for Theorem 3

Theorem 3. *Under the multi-hierarchical semantic consistency condition, the mutual information between the consensus representation $\mathbf{P}$ and the ground-truth labels $\mathbf{Y}$ is lower-bounded by the maximum mutual information achieved by any single view.*

$$I(\mathbf{P}; \mathbf{Y}) \geq \max_v I(\mathbf{P}^{(v)}; \mathbf{Y}) - \epsilon, \tag{29}$$

where $\epsilon \geq 0$ is a small constant representing the residual inconsistency across views.

Proof. From the data processing inequality, we have for any view v:

$$I(\mathbf{P}^{(v)}; \mathbf{Y}) \leq I(\mathbf{P}^{(v)}; \mathbf{P}; \mathbf{Y}) + I(\mathbf{P}^{(v)}; \mathbf{Y} | \mathbf{P}). \tag{30}$$

The consensus representation $\mathbf{P}$ is constructed by aggregating information from all view-specific representations $\{\mathbf{P}^{(v)}\}$. Under the multi-hierarchical semantic consistency condition: VCA ensures that $I(\mathbf{P}^{(v)}; \mathbf{P})$ is high, meaning $\mathbf{P}$ captures the information from each $\mathbf{P}^{(v)}$; CSR ensures that $I(\mathbf{P}; \mathbf{Y})$ is high, meaning $\mathbf{P}$ itself is discriminative.

Therefore, the conditional mutual information $I(\mathbf{P}^{(v)}; \mathbf{Y} | \mathbf{P})$ is small, as most information about $\mathbf{Y}$ in $\mathbf{P}^{(v)}$ is already captured in $\mathbf{P}$. This implies:

$$I(\mathbf{P}^{(v)}; \mathbf{Y}) \lesssim I(\mathbf{P}; \mathbf{Y}). \tag{31}$$

Taking the maximum over v on the left side of the inequality (31) and accounting for potential information loss during fusion (bounded by ϵ), we obtain the final result:

$$I(\mathbf{P}; \mathbf{Y}) \geq \max_v I(\mathbf{P}^{(v)}; \mathbf{Y}) - \epsilon. \tag{32}$$

□

A5. Complexity Analysis

We analyze the computational complexity of MCSF and compare it with conventional deep spectral clustering frameworks such as MvSCN [5], which rely on Siamese architectures. The analysis assumes mini-batch training, where n is the total number of samples and b is the batch size. The complexity of MCSF arises primarily from three components: (1) view-specific encoder-decoder networks, (2) spectral embedding with Cholesky orthogonalization, and (3) multi-hierarchical contrastive learning.

For each view, the encoder-decoder module has a complexity of $\mathcal{O}(b \sum d_i d_j)$, leading to a total of $\mathcal{O}(Vb \sum d_i d_j)$ across all V views. The batch-wise spectral embedding computation and Cholesky decomposition involve computing similarity graphs ($\mathcal{O}(Vb^2 d)$), orthogonalization ($\mathcal{O}(VK^3 + VbK^2)$), and contrastive loss over pairwise similarities ($\mathcal{O}(Vb^2 d + Vb^2)$). As a result, the per-epoch complexity of MCSF is:

$$\mathcal{O}\left(\frac{n}{b} \cdot V \cdot \left(2b^2 d + b^2 + K^3 + bK^2 + b \sum d_i d_j \right) \right). \tag{33}$$

Compared to frameworks that employ Siamese architectures for deep spectral embedding, MCSF significantly reduces computational overhead by eliminating redundant dual-branch encoders. For example, in MvSCN, each sample requires two forward passes through identical encoders, resulting in an additional complexity of $\mathcal{O}(2bF)$, where F denotes the cost of a single forward pass. In contrast, MCSF performs only a single pass per sample, reducing the feature

Algorithm 1: MCSF

Input: Multi-view data $\{\mathbf{X}^{(v)}\}_{v=1}^{V}$, cluster number K, hyperparameters α, β, neighborhood size k, batch size b; maximum epochs T

Output: Cluster labels

Initialize networks $\{h_\theta^{(v)}, h_\delta^{(v)}\}_{v=1}^{V}$;
for *epoch* $= 1$ **to** T **do**
  foreach *batch in dataloader* **do**
    for $v = 1$ **to** V **do**
      Encoder: $\mathbf{Z}^{(v)} \leftarrow h_\theta^{(v)}(\mathbf{X}^{(v)})$;
      Decoder: $\hat{\mathbf{X}}^{(v)} \leftarrow h_\delta^{(v)}(\mathbf{Z}^{(v)})$;
      Cluster probabilities: $\mathbf{H}^{(v)} \leftarrow \text{softmax}(g_\psi(\mathbf{Z}^{(v)}))$;
      Orthogonal embedding: $\mathbf{P}^{(v)} \leftarrow \mathbf{H}^{(v)}(\text{chol}((\mathbf{H}^{(v)})^\top \mathbf{H}^{(v)} + \varepsilon \mathbf{I}))^{-1}$;
      Compute similarity $\mathbf{W}^{(v)}$ via Eq. (1);
    end
    Initialize consensus representation: $\mathbf{P} \leftarrow \frac{1}{V}\sum_{v=1}^{V} \mathbf{P}^{(v)}$;
    Compute $\mathcal{L}_{\text{re}}$ via Eq. (14);
    Compute $\mathcal{L}_{\text{spc}}$ via Eq. (4);
    Compute $\mathcal{L}_c$ via Eq. (10);
    Total loss: $\mathcal{L} \leftarrow \mathcal{L}_{\text{re}} + \alpha\mathcal{L}_{\text{spc}} + \beta\mathcal{L}_c$;
    Update parameters via backpropagation;
  end
end
Apply k-means to $\mathbf{P}$ to obtain cluster labels.

extraction cost to $\mathcal{O}(bF)$. This design achieves approximately a twofold reduction in encoder-related computation per batch, enabling more efficient training without sacrificing clustering performance.

B. Experimental Details

B1. Algorithmic Procedure

The complete procedure of MCSF is summarized in Algorithm 1.

B2. Datasets

We evaluate MCSF on eight widely-used multi-view datasets that cover a range of domains, including text, objects, digits, and faces. 3Sources [12]: A text classification dataset containing 169 news articles from three sources, *i.e.*, BBC, Reuters, and The Guardian. MSRC-v1 [13]: An object recognition dataset with 210 images from 7 categories, each described by 5 visual feature representations. COIL-20 [8]: A classic multi-view object dataset comprising 1440 grayscale images of 20 objects, with three views extracted using different visual descriptors. Extended YaleB [4]: A face dataset with 640 images of 10 individuals captured under varying illumination conditions. MNIST-USPS [1]: A digit dataset combining the MNIST and USPS datasets, where each sample is converted into a 768-dimensional vector, consisting of 5000 samples and 10 clusters. Hdigit [2]: A large-scale handwritten digit dataset containing 10,000 samples. It provides two distinct feature views with dimensions 784 and 256, respectively. NUS-WIDE [3]: A subset of the real-world web image dataset containing 1600 images from 8 categories, described by six heterogeneous visual feature views. Cifar100 [7]: A dataset for object recognition consisting of 50,000 samples, 100 clusters, and 3 views.

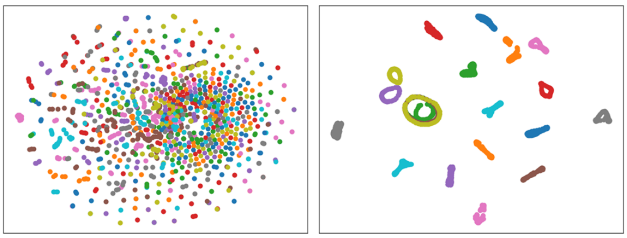

Figure 7. Visualization of the representation matrix ($\mathbf{Z}^{(1)}$) on COIL-20 without (left) and with (right) spectral embedding loss ($\mathcal{L}_{spc}$). Different colors denote different clusters.

Table 4. Batch size configurations on different datasets.

Datasets	b_1	b_2	b_3	b_4	b_5	b_6
EYaleB	64	80	128	160	320	640
COIL-20	48	144	240	480	720	1440
NUS-WIDE	50	100	200	400	800	1600
Hdigit	100	200	500	1000	2000	5000

C. Additional Experiments

C1. Experiments on Spectral Embedding Loss

In addition to the quantitative results, we further visualize the representation matrices learned with and without the spectral embedding loss ($\mathcal{L}_{spc}$) using t-SNE [11] on the COIL-20 dataset, as shown in Figure 7. When $\mathcal{L}_{spc}$ is removed (left), the learned representations are poorly separated and heavily entangled, leading to ambiguous cluster boundaries. In contrast, when $\mathcal{L}_{spc}$ is incorporated (right), the resulting representations exhibit well-separated and compact clusters, aligning with class semantics. This visual evidence reinforces the numerical findings in Table 1, particularly the substantial improvement on COIL-20 (from 56.46% to 84.86%) when adding $\mathcal{L}_{spc}$ to a model trained only with contrastive loss. It demonstrates that spectral embedding loss plays a critical role in uncovering the intrinsic geometric structure, thus enhancing inter-cluster separability and facilitating more effective clustering.

C2. Batch Size Configurations

The details of the batch size configurations corresponding to Figure 6 in the main paper are provided in Table 4.

References

[1] Arthur Asuncion, David Newman, et al. UCI machine learning repository, 2007. 3

[2] Man-Sheng Chen, Chang-Dong Wang, and Jian-Huang Lai. Low-rank tensor based proximity learning for multi-view clustering. *IEEE Trans. Knowl. Data Eng.*, 2022. 3

[3] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: a real-world web image database from National University of Singapore. In *ACM Int. Conf. Image Video Retr.*, pages 1–9, 2009. 3

[4] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):643–660, 2001. 3

[5] Zhenyu Huang, Joey Tianyi Zhou, Xi Peng, Changqing Zhang, Hongyuan Zhu, and Jiancheng Lv. Multi-view spectral clustering network. In *Int. Joint Conf. Artif. Intell.*, page 4, 2019. 1, 2

[6] Zhenyu Huang, Joey Tianyi Zhou, Hongyuan Zhu, Changqing Zhang, Jiancheng Lv, and Xi Peng. Deep spectral representation learning from multi-view data. *IEEE Trans. Image Process.*, 30:5352–5362, 2021. 1

[7] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3

[8] Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (COIL-20). 1996. 3

[9] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. 2017. 2

[10] Uri Shaham, Kelly Stanton, Henry Li, Boaz Nadler, Ronen Basri, and Yuval Kluger. Spectralnet: Spectral clustering using deep neural networks. In *Int. Conf. Learn. Represent.*, 2018. 1

[11] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *J. Mach. Learn. Res.*, 9(11), 2008. 3

[12] Liang Yao and Gui-Fu Lu. Double structure scaled simplex representation for multi-view subspace clustering. *Neural Netw.*, 151:168–177, 2022. 3

[13] Pei Zhang, Xinwang Liu, Jian Xiong, Sihang Zhou, Wentao Zhao, En Zhu, and Zhiping Cai. Consensus one-step multi-view subspace clustering. *IEEE Trans. Knowl. Data Eng.*, 34 (10):4676–4689, 2020. 3