Self-Enhanced Image Clustering with Cross-Modal Semantic Consistency

Zihan Li[1], Wei Sun[1], Jing Hu[2], Jianhua Yin[1], Xing Wang[3], Erwei Yin[4], Jianlong Wu[5,*]
[1]School of Computer Science and Technology, Shandong University,
[2]College of Civil Engineering, Tongji University,
[3]School of Information Science and Engineering, Linyi University,
[4]Defense Innovation Institute, Academy of Military Sciences,
[5]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
zihanli@mail.sdu.edu.cn, sunweimy@gmail.com, jinghu@utexas.edu, jhyin@sdu.edu.cn, wangying@lyu.edu.cn,
yinerwei1985@gmail.com, wujianlong@hit.edu.cn

Abstract

While large language-image pre-trained models like CLIP offer powerful generic features for image clustering, existing methods typically freeze the encoder. This creates a fundamental mismatch between the model's task-agnostic representations and the demands of a specific clustering task, imposing a ceiling on performance. To break this ceiling, we propose a self-enhanced framework based on cross-modal semantic consistency for efficient image clustering. Our framework first builds a strong foundation via cross-modal semantic consistency and then specializes the encoder through self-enhancement. In the first stage, we focus on cross-modal semantic consistency. By mining consistency between generated image-text pairs at the instance, cluster assignment, and cluster center levels, we train lightweight clustering heads to align with the rich semantics of the pre-trained model. This alignment process is bolstered by a novel method for generating higher-quality cluster centers and a dynamic balancing regularizer to ensure well-distributed assignments. In the second stage, we introduce a self-enhanced fine-tuning strategy. The well-aligned model from the first stage acts as a reliable pseudo-label generator. These self-generated supervisory signals are then used to feed back the efficient, joint optimization of the vision encoder and clustering heads, unlocking their full potential. Extensive experiments on six datasets show that our method outperforms existing deep clustering methods by significant margins. Notably, our results under ViT-B/32 model match or even surpass the accuracy of state-of-the-art methods built upon the far larger ViT-L/14.

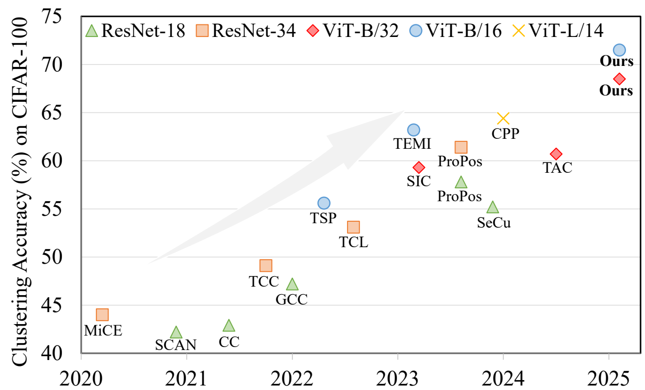


Figure 1. Comparison of existing deep clustering methods. The deep clustering paradigm gradually shifts from CNN-based to ViT-based with a steady increase in performance.

Introduction

Image clustering is a classic unsupervised task that aims to group unlabeled samples into different clusters by exploiting the inherent relationships between the samples. Initially, researchers mainly focused on traditional machine learning methods for image clustering, including K-means (Neyman and Scott 1967), hierarchical clustering (Ward Jr 1963), spectral clustering (Ng, Jordan, and Weiss 2001; Zelnik-Manor and Perona 2004), subspace clustering (Kailing, Kriegel, and Kröger 2004; Liu et al. 2019; Zhang et al. 2015, 2018; Xie et al. 2024), and concept factorization (Zhang et al. 2019; Cai et al. 2009), most of which relied heavily on prior assumptions about the data distribution. However, their performance decreases significantly when faced with complex and high-dimensional image data. With the emergence of deep learning, researchers began to incorporate deep learning into clustering objectives. Based on the powerful feature extraction capabilities, deep learning image clustering approaches can automatically and accurately capture the structure and semantics within the data, exhibiting outstanding clustering performance (Tsai, Li, and Zhu 2020; Vincent et al. 2010; Radford, Metz, and Chintala 2016; Zeiler et al. 2010; Kingma and Welling 2014; Xie, Girshick, and Farhadi 2016; Chang et al. 2017; Wu et al. 2019; Qian 2023; Metaxas, Tzimiropoulos, and Patras 2023; Wu et al. 2025; Wang et al. 2025). With the development of Vision Transformers (ViT) (Dosovitskiy et al. 2021), the paradigm of deep clustering is gradually shifting from traditional CNN-based methods to ViT-based approaches, as illustrated in Fig. 1. Compared with convolutional neural networks that capture local features through sliding convolutional kernels, ViTs employ a self-attention mechanism, which can effectively capture the global semantic informa-

[*]Corresponding author.

tion, thus obtaining higher-quality features and improving clustering results. Due to the huge amount of parameters in ViT, training ViT from scratch in an end-to-end manner requires significant computational resources. As a result, existing ViT-based clustering methods, such as TSP (Zhou and Zhang 2022), TEMI (Adaloglou et al. 2023), CPP (Chu et al. 2024), etc., typically rely on pre-trained vision encoders that are completely frozen during training and introduce a trainable projection head and clustering head to simultaneously learn task-specific features and cluster assignments. Considering the limitation that relying solely on visual features for training may incorrectly cluster visually similar but semantically dissimilar samples together, recent methods like SIC (Cai et al. 2023) and TAC (Li et al. 2024) have further incorporated additional textual information from the WordNet (Miller 1995) dictionary to assist in image clustering tasks, successfully enhancing clustering performance. However, this frozen-encoder approach introduces a fundamental bottleneck: it creates a mismatch between the general-purpose, task-agnostic features of the pre-trained model and the highly specific, discriminative representations needed for the target clustering task. This inevitably caps the model's performance, regardless of how well the attached clustering heads are trained. The key, therefore, is to find a way to safely unlock and adapt the encoder to the specific task without supervised labels.

To address this limitation, we propose a novel two-stage framework, **S**elf-**E**nhanced **I**mage **C**lustering (SEIC), designed to first align with the pre-trained knowledge and then specialize the encoder for the task at hand. Our first step is to build a strong foundational model without altering the powerful pre-trained encoder. We start by generating image-text feature pairs using knowledge from CLIP and WordNet. Then, we train the clustering heads by enforcing Cross-Modal Semantic Consistency at three distinct levels: instance features, cluster assignments, and cluster centers. This multi-level alignment forces the heads to learn representations that are consistent with the rich semantics of the CLIP space. To strengthen this alignment, we introduce a novel method for computing higher-quality cluster centers based on assignment probabilities and a dynamic balance regularizer that adapts to the sample learning status to mitigate cluster collapse.Once the heads are well-aligned, the model can produce reliable, high-confidence pseudo-labels. In the second stage, we leverage this capability for Self-Enhancement. The high-quality pseudo-labels serve as self-generated supervisory signals that feed back to guide the fine-tuning process. Using efficient parallel LoRA adapters (Hu et al. 2021), we jointly optimize the vision encoder and the clustering heads. This critical step allows the encoder to move beyond its generic initialization and specialize its feature extraction capabilities for the specific data distribution, thereby breaking the performance ceiling imposed by the frozen-encoder paradigm.

Overall, our contributions can be summarized as follows:

- We propose a novel two-stage **"align, then enhance"** framework (SEIC) that resolves the feature mismatch problem in pre-trained models by first building a strong foundational model and then specializing the encoder for clustering.
- For the alignment stage, we enforce Cross-Modal Semantic Consistency at multiple levels, supported by a novel probability-weighted cluster center generation method and a dynamic balancing regularizer.
- For the enhancement stage, we introduce a Self-Enhanced Efficient Fine-tuning where the model uses its own high-confidence predictions to feed back and efficiently refine the vision encoder via LoRA.
- Extensive experiments on six benchmark datasets demonstrate that SEIC significantly outperforms existing state-of-the-art methods, even with smaller backbones.

Related Work

Pre-trained ViT Models

The success of deep clustering is increasingly tied to powerful pre-trained models. Self-supervised methods like MoCo-v3 (Chen, Xie, and He 2021), MAE (He et al. 2022), and DINO (Caron et al. 2021) learn robust visual representations from images alone. More recently, vision-language models such as CLIP (Radford et al. 2021) have shown remarkable zero-shot capabilities by learning from massive image-text pairs. Our work investigates how to effectively adapt the rich knowledge from these models, particularly CLIP, for the unsupervised clustering task.

However, the research on utilizing the knowledge of pre-trained models to improve the image clustering task is still in the exploratory stage. Therefore, in this paper, we propose a self-enhanced framework based on the pre-trained model with both image and text modalities to provide more insights and solutions for this issue.

Deep Image Clustering

In the past, deep image clustering has primarily relied on ResNet (He et al. 2016) as the backbone, enhancing its performance through various optimization techniques. Recently, the introduction of contrastive learning has significantly advanced deep image clustering, with methods such as (Li et al. 2021b; Zhong et al. 2021; Shen et al. 2021; Li et al. 2022; Van Gansbeke et al. 2020; Li et al. 2021a; Huang et al. 2022; Niu, Shan, and Wang 2022; Metaxas, Tzimiropoulos, and Patras 2023; Qian 2023). Among them, CC (Li et al. 2021b), GCC (Zhong et al. 2021), and TCL (Li et al. 2022) utilize contrastive learning at both the instance and cluster levels. CC and TCL create positive and negative samples through data augmentation, while GCC expands the selection of such samples through K-nearest neighbors. PCL (Li et al. 2021a) and ProPos (Huang et al. 2022) focus solely on instance comparison, utilizing prototype contrast to enrich the selection of positive samples. Some methods based on pseudo-label have also been proposed to further improve clustering model performance, such as SCAN (Van Gansbeke et al. 2020) and SPICE (Niu, Shan, and Wang 2022). SCAN introduces the self-labeling method, further enhancing model performance by selecting high-confidence pseudo-labels from a trained clustering model. SPICE employs semi-supervised training by fixing a portion of pseudo-labels for model fine-tuning.

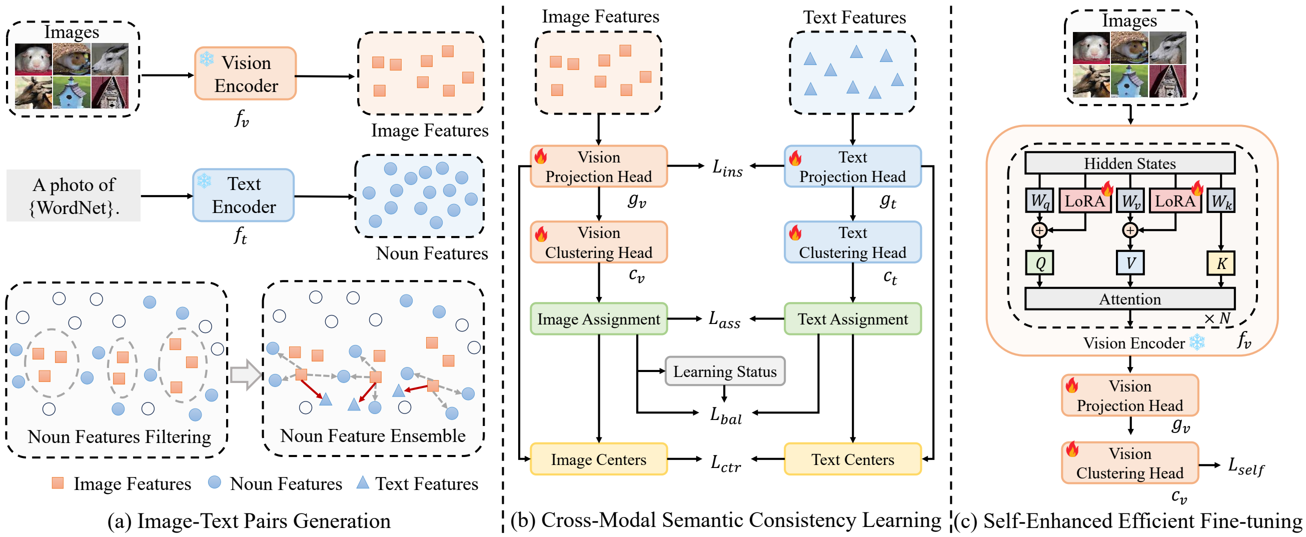


Figure 2. The framework of our proposed SEIC method. (a) We first construct image-text feature pairs using the pre-trained model CLIP and the dictionary WordNet. (b) Then, we perform cross-modal semantic consistency learning to achieve more discriminative representations and accurate clustering assignments. (c) Finally, a self-enhanced fine-tuning strategy is proposed to further improve the clustering performance.

Recently, several deep image clustering methods have sought to advance clustering tasks by utilizing pre-trained ViT models. TEMI (Adaloglou et al. 2023) employs DINO (Caron et al. 2021) to reduce the distance between images and their neighbors. TSP (Zhou and Zhang 2022) undergoes training through self-distillation and is trained on the necessary pre-trained model. SIC (Cai et al. 2023) and TAC (Li et al. 2024) predominantly use the CLIP (Radford et al. 2021) pre-trained model, effectively leveraging its characteristics by constructing text features. However, all of these models freeze the encoders and train only the appended heads, limiting the model to task-agnostic features. In our approach, we learn discriminative representations and correct assignments by mining multi-level cross-modal semantic consistency and propose a self-enhanced fine-tuning strategy to jointly fine-tune encoder and head components.

Parameter-Efficient Fine-Tuning

Fine-tuning large models is computationally expensive. PEFT methods, such as adapters (Houlsby et al. 2019) and Low-Rank Adaptation (LoRA) (Hu et al. 2021), address this by updating only a small number of parameters. While these techniques are well-established in supervised learning, their application in a fully unsupervised clustering scenario remains less explored. Our work presents an effective way to leverage LoRA for unsupervised encoder specialization.

Method

Problem Formulation and Overview of SEIC

Given N unlabeled images $\mathcal{I} = [I_1, ..., I_N]$, deep clustering aims to group them into K different clusters, ensuring that images of the same semantic labels are grouped together as much as possible. As shown in Fig. 2, we design a self-enhanced framework SEIC for image clustering based on image-text pairs with the help of the pre-trained model CLIP (Radford et al. 2021). The model consists of vision and text pre-trained encoders and corresponding head components. Specifically, the vision encoder f_v and text encoder f_t are able to transform the images as well as the corresponding texts into features, leveraging the knowledge of CLIP. The vision projection head g_v and text projection head g_t aim to project features into a space more suitable for clustering. The clustering heads c_v and c_t are used for cluster assignment, i.e., mapping the projected image or text feature to a K-dimensional probability vector p_i that satisfies $\sum_{j=1}^{K} p_{ij} = 1$, where p_{ij} denotes the probability that the image I_i is assigned to the j-th cluster. Finally, $r_i = \arg\max(p_i)$ is the clustering result for image I_i.

The training process of SEIC includes two stages:
1) **Alignment via Cross-Modal Semantic Consistency**: We first freeze the CLIP encoders and train only the head components. The goal is to align the heads' behavior with the rich, latent semantic space of CLIP, thereby creating a strong foundational clustering model.
2) **Self-Enhancement for Encoder Specialization**: With a well-aligned model, we then proceed to fine-tune the vision encoder f_v itself. The model uses its own high-confidence predictions to generate supervisory signals, which *feed back* to specialize the encoder for the target dataset.

Alignment via Cross-Modal Semantic Consistency

In this stage, our objective is to learn high-quality projection and clustering heads by maximally extracting knowledge from the frozen CLIP encoders. This process involves generating cross-modal data and then enforcing consistency at multiple semantic levels.

Image-Text Pair Generation In the deep clustering task, where image samples lack textual annotations, our goal is to construct image-text pairs $\mathcal{S} = \{v_i, t_i\}_{i=1}^N$. Based on the premise that nouns closer to an image in the shared feature space better reflect its content, we utilize the WordNet dictionary to generate textual information for each image.

Similar to TAC (Li et al. 2024), specifically, we first extract image features $\mathcal{V} = \{v_i\}_{i=1}^N$ using the frozen vision encoder f_v. All nouns from WordNet are encoded into features $\mathcal{N}$ using the text encoder f_t. To manage computational costs, we create a candidate noun subset $\overline{\mathcal{N}}$ by selecting the top k_1 nearest noun features for each of the K initial image cluster centers (obtained via K-means). Finally, for each image feature v_i, its corresponding text feature t_i is constructed as a weighted sum of its top-k_2 most similar noun features from $\overline{\mathcal{N}}$:

$$t_i = \sum_{j=1}^{|\overline{\mathcal{N}}|} \hat{s}_{ij} \overline{n}_j, \tag{1}$$

where $\hat{s}_{ij}$ is non-zero only if $\overline{n}_j$ is one of the top-k_2 neighbors of v_i. This provides us with the image-text pairs $\{v_i, t_i\}$ needed for alignment.

Cross-Modal Semantic Consistency Learning With the image-text pairs, we now align the heads by enforcing consistency at the instance feature, cluster assignment, and cluster center levels.

For instance feature level consistency, we treat the image and text projected features obtained from the projection heads as positive pairs for contrastive learning. To be specific, we first compute the projected features $\widetilde{v}_i = g_v(v_i), \widetilde{t}_i = g_t(t_i)$, based on which we construct the bidirectional contrastive loss as follows:

$$L_{ins}^{(v \to t)} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(\langle \widetilde{v}_i, \widetilde{t}_i \rangle / \tau)}{\sum_{j=1}^{B} \exp(\langle \widetilde{v}_i, \widetilde{t}_j \rangle / \tau)}, \tag{2}$$

$$L_{ins} = (L_{ins}^{(v \to t)} + L_{ins}^{(t \to v)})/2, \tag{3}$$

where τ is a trainable scaling factor and B is the batch size. L_{ins} is structurally similar to the loss proposed in CLIP, which allows the projection heads to effectively align with the distribution of image-text features learned by CLIP, thereby generating more clustering-friendly features.

At the cluster assignment level, the clustering results from the image and text clustering heads should also share consistent distributions. Therefore, we first achieve the cluster assignment probabilities under different modalities based on the projected feature, i.e., $p_i^v = c_v(\widetilde{v}_i), p_i^t = c_t(\widetilde{t}_i)$. For each column of p^v and p^t within a mini-batch, we treat them as a positive pair and perform contrastive learning as follows:

$$L_{ass}^{(v \to t)} = -\frac{1}{K} \sum_{i=1}^{K} \log \frac{\exp(\langle \ddot{p}_i^v, \ddot{p}_i^t \rangle / \hat{\tau})}{\sum_{j=1}^{K} \exp(\langle \ddot{p}_i^v, \ddot{p}_j^t \rangle / \hat{\tau})}, \tag{4}$$

$$L_{ass} = (L_{ass}^{(v \to t)} + L_{ass}^{(t \to v)})/2, \tag{5}$$

where $\ddot{p}_i^v$ represents the i-th column of p^v, and $\hat{\tau}$ is a fixed temperature coefficient. This loss encourages the clustering heads to produce more accurate cluster assignments.

Considering that we only explore feature consistency at the instance level, which ignores potential category information between samples, we further extend our approach to incorporate cluster center-level consistency. It ensures that the centers sharing the same category across different modalities are maximally proximate in feature space, while those of different categories are significantly distanced apart. To achieve more accurate cluster centers, we propose a novel generation method. Unlike the conventional approach of directly taking the mean of all sample features within a cluster as the center (Li et al. 2021a; Huang et al. 2022), we construct the center based on the contribution of samples to this cluster, which is reflected by clustering predicted probability. Specifically, each cluster center is formed by the weighted sum of projection features from samples belonging to this center based on their contribution degree, calculated by:

$$\mu_k^v = \sum_{i=1}^{B} \mathbb{1}(\arg\max(p_i^v) = k) \cdot \bar{p}_{ik}^v \cdot \widetilde{v}_i, \tag{6}$$

$$\mu_k^t = \sum_{i=1}^{B} \mathbb{1}(\arg\max(p_i^t) = k) \cdot \bar{p}_{ik}^t \cdot \widetilde{t}_i, \tag{7}$$

where $\mathbb{1}$ is the indicator function and $\bar{p}_i$ is the L1-normalized version of the probability vector p_i. We then treat the cluster centers of the same category under the image and text modalities as a positive pair and bring them closer in the feature space:

$$L_{ctr}^{(v \to t)} = -\frac{1}{K} \sum_{i=1}^{K} \log \frac{\exp(\langle \mu_i^v, \mu_i^t \rangle / \hat{\tau})}{\sum_{j=1}^{K} \exp(\langle \mu_i^v, \mu_j^t \rangle / \hat{\tau})}. \tag{8}$$

$$L_{ctr} = (L_{ctr}^{(v \to t)} + L_{ctr}^{(t \to v)})/2. \tag{9}$$

Compared with the traditional PCL (Li et al. 2021a) and ProPos (Huang et al. 2022) methods, our proposed construction method can weigh samples within a batch based on their proximity to these centers. We facilitate the emergence of high-quality cluster centers.

Stabilizing Alignment with Dynamic Balancing To prevent the trivial solution of assigning all samples to a few clusters, we introduce a dynamic balancing regularizer. Instead of a static entropy loss, our term adapts to the model's learning state. We track the historical assignment distribution h for the image modality via an exponential moving average (EMA): $h \leftarrow m \cdot h + (1 - m) \cdot \text{Hist}(\arg\max(p^v))$. The balance loss is then defined as:

$$L_{bal} = \sum_{j=1}^{K} \frac{b_j^v \log(b_j^v) + b_j^t \log(b_j^t)}{h_j}, \tag{10}$$

where b_j^m is the average predicted probability for cluster j in the current batch for modality $m \in \{v, t\}$. This dynamically up-weights the penalty for under-populated clusters, encouraging a more uniform assignment.

The final objective for the alignment stage is a weighted sum of these components:

$$L_{align} = \alpha L_{ins} + \beta L_{ass} + \gamma L_{ctr} + \delta L_{bal}. \tag{11}$$

Dataset		CIFAR-10			CIFAR-100			STL-10			ImageNet-10			ImageNet-Dogs			Tiny-ImageNet		
Method	Backbone	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
K-means	-	0.087	0.229	0.049	0.084	0.130	0.028	0.125	0.192	0.061	0.119	0.241	0.057	0.055	0.105	0.020	0.065	0.025	0.005
SCAN	ResNet-18	0.797	0.883	0.772	0.486	0.507	0.333	0.698	0.809	0.646	-	-	-	-	-	-	-	-	-
GCC	ResNet-18	0.764	0.856	0.728	0.472	0.472	0.305	0.684	0.788	0.631	0.842	0.901	0.822	0.490	0.526	0.322	0.347	0.138	0.075
ProPos	ResNet-18	0.853	0.921	0.844	0.572	0.573	0.417	-	-	-	-	-	-	-	-	-	0.405	0.256	0.143
ProPos	ResNet-34	0.886	0.943	0.884	0.606	0.614	0.451	0.758	0.867	0.737	0.896	0.956	0.906	0.692	0.745	0.627	-	-	-
TCL	ResNet-34	0.819	0.887	0.780	0.529	0.531	0.357	0.799	0.868	0.757	0.875	0.895	0.837	0.623	0.644	0.516	-	-	-
TCC	ResNet-34	0.790	0.906	0.733	0.479	0.491	0.312	0.732	0.814	0.689	0.848	0.897	0.825	0.554	0.595	0.417	-	-	-
TSP	DINO ViT-S/16	0.847	0.921	0.838	0.582	0.549	0.408	0.941	0.970	0.938	-	-	-	-	-	-	-	-	-
TSP	DINO ViT-B/16	0.880	0.940	0.875	0.614	0.556	0.433	0.958	0.979	0.956	-	-	-	-	-	-	-	-	-
SIC	CLIP ViT-B/32	0.847	0.926	0.844	0.593	0.583	0.439	0.953	0.981	0.959	0.970	0.982	0.961	0.690	0.697	0.558	-	-	-
TEMI	DINO ViT-B/16	0.886	0.945	0.885	0.654	0.632	0.489	0.965	0.985	0.968	-	-	-	-	-	-	-	-	-
TEMI	CLIP ViT-L/14	0.926	0.969	0.932	0.645	0.618	0.468	0.964	0.974	0.949	-	-	-	-	-	-	-	-	-
CPP	CLIP ViT-L/14	0.936	0.974	-	0.725	0.642	-	-	-	-	-	-	-	-	-	-	-	-	-
TAC	CLIP ViT-B/32	0.841	0.923	0.839	0.611	0.607	0.448	0.955	0.982	0.961	0.985	0.992	0.983	0.806	0.830	0.722	-	-	-
CLIP	CLIP ViT-B/32	0.703	0.742	0.616	0.499	0.455	0.283	0.917	0.943	0.891	0.969	0.982	0.961	0.398	0.381	0.201	0.543	0.373	0.219
SEIC (Ours)	CLIP ViT-B/32	0.938	0.976	0.947	0.691	0.685	0.554	0.963	0.987	0.970	0.988	0.996	0.991	0.846	0.889	0.817	0.701	0.598	0.472
SEIC (Ours)	CLIP ViT-B/16	**0.945**	**0.978**	**0.952**	**0.732**	**0.715**	**0.598**	**0.980**	**0.993**	**0.985**	**0.992**	**0.997**	**0.994**	**0.886**	**0.910**	**0.858**	**0.730**	**0.644**	**0.523**

Table 1. Clustering results of various methods on six widely-used datasets. The best and second-best results are shown in **bold** and underline, respectively.

Self-Enhanced Efficient Fine-tuning

The alignment stage yields high-quality clustering heads, but the vision encoder remains generic. To break this performance ceiling, we now unfreeze the vision encoder and specialize it for the target dataset.

A naive approach would be to continue training with L_{align}. However, this fails because the text features $\{t_i\}$ are static, pre-computed by the frozen text encoder. As the vision encoder f_v updates, its features v_i will drift, breaking the semantic correspondence with the fixed t_i and leading to performance collapse. We need a new supervisory signal that relies only on the visual modality.

The well-aligned model from Stage 1 is now a confident "teacher" capable of generating reliable pseudo-labels for itself. We use these pseudo-labels to create a self-supervision loss. To handle potential noise in the pseudo-labels, we adopt a confidence-based weighting strategy inspired by SoftMatch (Chen et al. 2023). A sample's contribution to the loss is weighted by its prediction confidence, measured by a truncated Gaussian function:

$$w_i = \begin{cases} \exp\left(-\frac{(\max(p_i^v)-\mu_t)^2}{2\sigma_t^2}\right), & \text{if } \max(p_i^v) < \mu_t, \\ 1, & \text{otherwise,} \end{cases} \tag{12}$$

where μ_t and σ_t^2 are the moving average of the mean and variance of the maximum prediction probabilities, respectively. The self-enhancement loss is then a weighted cross-entropy:

$$L_{self} = \frac{1}{B}\sum_{i=1}^{B} w_i \cdot H(q_i^v, \arg\max(p_i^v)), \tag{13}$$

where H is the cross-entropy loss, q_i^v is the prediction for an augmented view of image I_i, and $\arg\max(p_i^v)$ is the pseudo-label from the original view.

To fine-tune the vision encoder efficiently and prevent catastrophic forgetting, we integrate lightweight Low-Rank Adaptation (LoRA) (Hu et al. 2021) adapters into the Q and V matrices of its self-attention blocks. The final optimization in this stage jointly trains the LoRA parameters in f_v along with the vision heads (g_v, c_v) using L_{self}. This self-enhancement process allows the encoder to learn task-specific features, leading to a significant performance boost.

Experiments

Compared Methods

We compared traditional clustering methods and deep image clustering methods using backbones of various sizes. The baseline methods include K-means (Neyman and Scott 1967), SCAN (Van Gansbeke et al. 2020), GCC (Zhong et al. 2021), ProPos (Huang et al. 2022), TCL (Li et al. 2022), TCC (Shen et al. 2021), TSP (Zhou and Zhang 2022), SIC (Cai et al. 2023), TEMI (Adaloglou et al. 2023), CPP (Chu et al. 2024), and TAC (Li et al. 2024), all of which utilize ResNet-18, ResNet-34, ViT-S/16, ViT-B/32, ViT-B/16 and ViT-L/14 as backbone. For CLIP, performance is evaluated by directly applying K-means clustering on the extracted image features.

Main Results

To validate the effectiveness of SEIC, we conducted extensive comparisons with state-of-the-art methods across all six datasets. The main results are presented in Table 1. Our method consistently outperforms existing approaches across all metrics and backbones. For instance, on CIFAR-100, our ViT-B/16 model achieves an ACC of 71.5%, surpassing the DINO-based TEMI (63.2%) and even the larger ViT-L/14-based CPP (64.2%) by a significant margin. This highlights the effectiveness of our "align, then enhance" strategy.

Table 2 further details the performance across different backbone sizes, demonstrating the consistent advantage of our two-stage approach. The full SEIC model substantially improves upon its Stage 1 variant (SEIC†), confirming the benefits of the self-enhancement stage. Notably, our SEIC† (Stage 1 only) is already competitive with, or superior to, strong baselines like TAC. In terms of efficiency, on an RTX 3090, the alignment stage for CIFAR-100 takes only 5 minutes, and the self-enhancement stage takes approximately 2 hours.

Method	Backbone	CIFAR-100	ImageNet-Dogs
SIC	CLIP ViT-B/32	0.583	0.697
TAC	CLIP ViT-B/32	0.607	0.830
SEIC (ours)†	CLIP ViT-B/32	0.608	0.833
SEIC (ours)	CLIP ViT-B/32	0.685	0.889
TAC	CLIP ViT-B/16	-	0.857
TEMI	DINO ViT-B/16	0.632	-
SEIC (ours)†	CLIP ViT-B/16	0.644	0.869
SEIC (ours)	CLIP ViT-B/16	0.715	0.910
TEMI	CLIP ViT-L/14	0.618	-
CPP	CLIP ViT-L/14	0.642	-
SEIC (ours)†	CLIP ViT-L/14	0.698	0.893

Table 2. The results of different backbones. Metric: ACC. † indicates that the self-enhanced efficient fine-tuning strategy is not used.

L_{ins}	L_{ass}	L_{ctr}	CIFAR-100	Tiny-ImageNet
			0.082	0.030
		✓	0.403	0.463
	✓		0.604	0.024
✓			0.190	0.138
	✓	✓	0.637	0.536
✓		✓	0.541	0.486
✓	✓		0.631	0.287
✓	✓	✓	0.644	0.561

Table 3. Effectiveness of individual components. Backbone: ViT-B/16. Metric: ACC.

Ablation Study

In this section, we dissect the contributions of the key components within our framework. Unless otherwise specified, these experiments are conducted using the Stage 1 model to isolate the effects of the alignment components.

Effectiveness of Individual Components As shown in Table 3, we evaluated the effect of different levels of semantic consistency on clustering performance. It can be observed that the removal of any component (especially for cluster assignment and the cluster center level) results in a significant performance decrease. In addition, different datasets exhibit varying gains from different losses. For example, omitting the center-level consistency loss leads to a 28% performance drop on Tiny-ImageNet while only a 1% drop on CIFAR-100, which demonstrates that the cluster center level consistency is crucial for performance improvement, particularly in scenarios with a high number of categories.

Effectiveness of Balance Regularization Term We also explored the effectiveness of our proposed new balance regularization term. The results are presented in Table 4. We can see that our designed regularization term can bring 7% performance improvement compared to the old one. Additionally, as shown in Fig. 4, we visually illustrated the dynamic changes in the standard deviation of the distribution of generated pseudo-labels during the training process to highlight that our proposed balance term could produce more uniformly distributed pseudo-labels, thus further mitigating the clustering collapse issues.

Method	Tiny-ImageNet
SEIC (ours) w/o balance term	0.390
SEIC (ours) w/ original balance term	0.490
SEIC (ours)	0.561

Table 4. Effectiveness of balance regularization term. Backbone: ViT-B/16. Metric: ACC.

Method	CIFAR-100	Tiny-ImageNet
Mean strategy	0.561	0.438
SEIC (ours)	0.644	0.561

Table 5. Effectiveness of center generation. Backbone: ViT-B/16. Metric: ACC.

Effectiveness of Center Generation In order to verify the effectiveness of our designed method of constructing cluster centers, we compared it with the traditional way, i.e., taking the mean of the sample features within the clusters as the center. The comparison results are shown in Table 5. In contrast to the traditional way, our proposed weight-based construction method shows significant advantages, especially on Tiny-ImageNet, where the gap reaches 12.3%, which strongly demonstrates that our method enables the generation of more precise and reasonable centers.

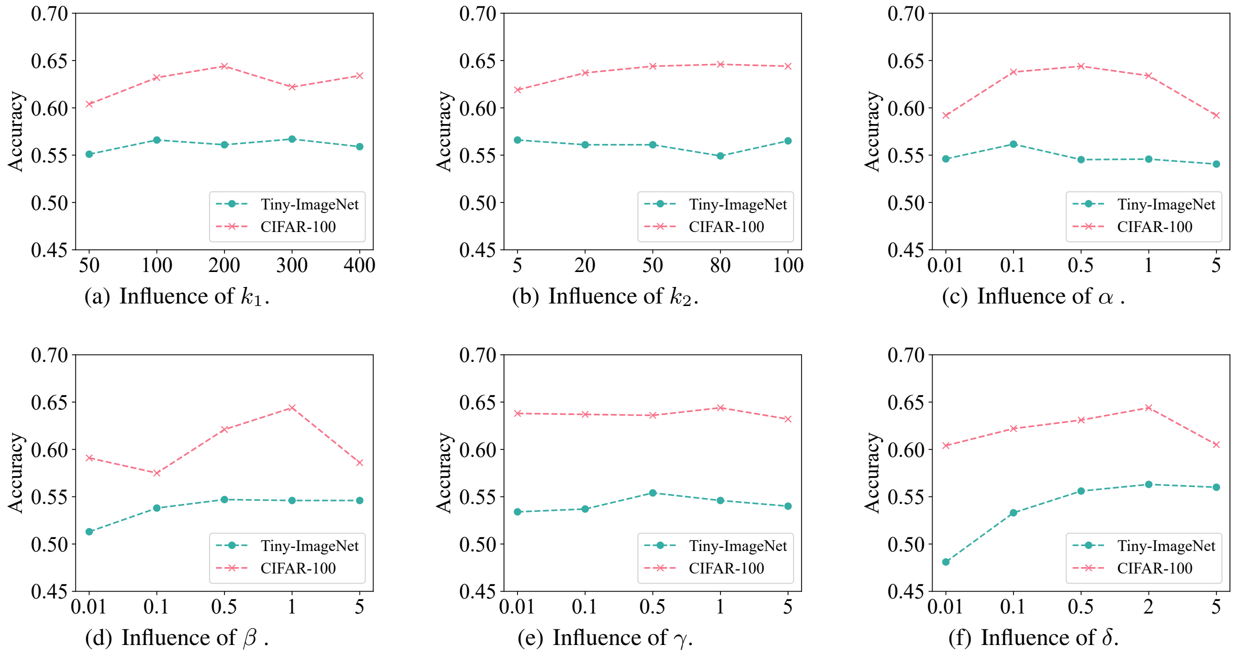
Parameter Sensitivity Analysis

Parameters in Image-Text Pairs Generation Stage We primarily focused on the ablation of k_1 and k_2, which significantly impact the quality of the generated text features. We can see from Fig. 3 (a) and (b) that setting k_1 and k_2 too large or small will reduce the clustering performance. Intuitively, when k_1 or k_2 is too small, there are too few noun features used to construct the text feature, resulting in less rich semantic information in the generated features. Conversely, if k_1 or k_2 is overlarge, it will introduce too many irrelevant nouns, thus lowering the quality of the generated text feature. We adopted a configuration that performs well across all datasets, with $k_1 = 200$ and $k_2 = 50$, ensuring the presence of both rich nouns and high-quality text features.

Loss Weights in Cross-Modal Semantic Consistency Learning Stage We also performed sensitivity analysis on the loss weights α, β, γ and δ, which are involved in cross-modal semantic consistency learning loss, to prove the robustness of our method, as shown in Fig. 3 (c), (d), (e), and (f), respectively. It can be found that the performance of our proposed method is stable on Tiny-ImageNet when all loss weights vary from 0.1 to 5. On the CIFAR-100, our method is also insensitive to all weights over specific intervals, such as α varying between $\{0.1, 0.5, 1\}$.

Analysis of the Self-Enhanced Efficient Fine-tuning

To better understand our proposed self-enhanced efficient fine-tuning strategy, we demonstrated the impact of different LoRA adapter settings on the clustering results and reveal the beneficial effects of different losses for self-enhanced learning on the clustering performance.



(a) Influence of k_1. (b) Influence of k_2. (c) Influence of α .

(d) Influence of β . (e) Influence of γ. (f) Influence of δ.

Figure 3. Parameter Sensitivity Analysis. Backbone: ViT-B/16. Metric: ACC.

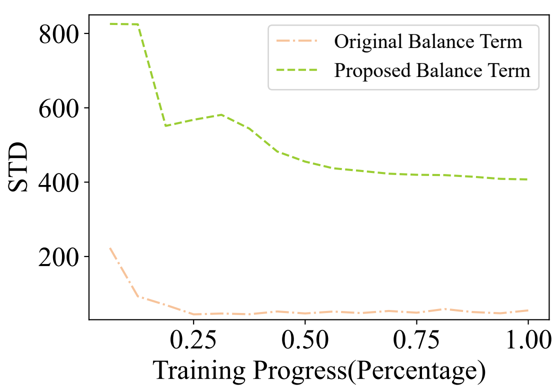


Figure 4. Dynamic changes in the standard deviation of the distribution of generated pseudo-labels during the training process. A lower standard deviation indicates a more balanced assignment distribution.

Self-Enhanced Learning Loss	CIFAR-100
Original Loss	0.247
Supervised CrossEntropy loss	0.664
FixMatch Loss	0.704
Self-Enhanced Loss	0.715

Table 6. Effectiveness of self-enhanced learning loss. Dataset: CIFAR-100. Backbone: ViT-B/16. Metric: ACC.

Self-Enhanced Learning Losses Table 6 compares different loss functions for the self-enhancement stage. As hypothesized in our method section, continuing to fine-tune with the original alignment loss (L_{align}) causes a performance collapse (24.7% ACC). This is because the vision encoder's features drift away from the static, pre-computed text features. In contrast, using a self-supervisory signal based on pseudo-labels is highly effective. Both FixMatch (Sohn et al. 2020) and our proposed confidence-weighted loss (L_{self}) yield strong results. Our L_{self} performs best, demonstrating the benefit of dynamically weighting samples based on prediction confidence.

Conclusion

In this paper, we introduced SEIC, a self-enhanced framework that significantly improves deep image clustering by addressing the critical feature mismatch problem inherent in frozen pre-trained models. Our two-stage "align, then enhance" approach first establishes a robust baseline by aligning clustering heads with CLIP's rich semantics through multi-level cross-modal consistency. This alignment is further strengthened by our novel probability-weighted cluster center generation mechanism and a dynamic balancing regularizer, which collectively enhance representation stability. Subsequently, the model iteratively feeds back its own high-confidence predictions to efficiently fine-tune the vision encoder, thereby adapting it more precisely to the specific clustering task. Extensive experiments validated that SEIC sets a new state-of-the-art on six challenging benchmarks, consistently and often by a significant margin. As a limitation, our reliance on explicit textual semantics may hinder performance in language-limited domains like medical imaging. Future work could explore generating pseudo-textual cues or developing non-language-based adaptation methods to extend our framework's applicability to a wider range of real-world scenarios.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62376069, Grant 92570108, Grant 62172261, and Grant 62236003, in part by the Young Elite Scientists Sponsorship Program by CAST under Grant 2023QNRC001, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515012027, in part by the Shenzhen Science and Technology Program under Grant KQTD20240729102207002 and Grant ZDSYS20230626091203008, and in part by Jiangsu Science and Technology Major Program under Grant BG2024041.

References

Adaloglou, N.; Michels, F.; Kalisch, H.; and Kollmann, M. 2023. Exploring the Limits of Deep Image Clustering using Pretrained Models. In *Proceedings of the British Machine Vision Conference*.

Cai, D.; He, X.; Wang, X.; Bao, H.; and Han, J. 2009. Locality preserving nonnegative matrix factorization. In *Proceedings of International Joint Conference on Artificial Intelligence*, 1010–1015.

Cai, S.; Qiu, L.; Chen, X.; Zhang, Q.; and Chen, L. 2023. Semantic-enhanced image clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 6869–6878.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, 9650–9660.

Chang, J.; Wang, L.; Meng, G.; Xiang, S.; and Pan, C. 2017. Deep adaptive image clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, 5879–5887.

Chen, H.; Tao, R.; Fan, Y.; Wang, Y.; Wang, J.; Schiele, B.; Xie, X.; Raj, B.; and Savvides, M. 2023. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. In *Proceedings of the International Conference on Learning Representations*.

Chen, X.; Xie, S.; and He, K. 2021. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, 9640–9649.

Chu, T.; Tong, S.; Ding, T.; Dai, X.; Haeffele, B. D.; Vidal, R.; and Ma, Y. 2024. Image clustering via the principle of rate reduction in the age of pretrained models. *Proceedings of the International Conference on Learning Representations*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 16000–16009.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the International Conference on Machine Learning*, 2790–2799. PMLR.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations*.

Huang, Z.; Chen, J.; Zhang, J.; and Shan, H. 2022. Learning representation for clustering via prototype scattering and positive sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kailing, K.; Kriegel, H.-P.; and Kröger, P. 2004. Density-connected subspace clustering for high-dimensional data. In *Proceedings of the International Conference on Data Mining*, 246–256.

Kingma, D. P.; and Welling, M. 2014. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations*.

Li, J.; Zhou, P.; Xiong, C.; and Hoi, S. C. 2021a. Prototypical contrastive learning of unsupervised representations. In *Proceedings of the International Conference on Learning Representations*.

Li, Y.; Hu, P.; Liu, Z.; Peng, D.; Zhou, J. T.; and Peng, X. 2021b. Contrastive clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 8547–8555.

Li, Y.; Hu, P.; Peng, D.; Lv, J.; Fan, J.; and Peng, X. 2024. Image clustering with external guidance. *Proceedings of the International Conference on Machine Learning*.

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.

Liu, G.; Zhang, Z.; Liu, Q.; and Xiong, H. 2019. Robust subspace clustering with compressed data. *IEEE Transactions on Image Processing*, 28(10): 5161–5170.

Metaxas, I. M.; Tzimiropoulos, G.; and Patras, I. 2023. Divclust: Controlling diversity in deep clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3418–3428.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41.

Neyman, J.; and Scott, E. L. 1967. Berkeley symposium on mathematical statistics and probability. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*.

Ng, A.; Jordan, M.; and Weiss, Y. 2001. On spectral clustering: Analysis and an algorithm. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 14.

Niu, C.; Shan, H.; and Wang, G. 2022. Spice: Semantic pseudo-labeling for image clustering. *IEEE Transactions on Image Processing*, 31: 7264–7278.

Qian, Q. 2023. Stable Cluster Discrimination for Deep Clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, 16645–16654.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International conference on machine learning*, 8748–8763. PMLR.

Radford, A.; Metz, L.; and Chintala, S. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *Proceedings of the International Conference on Learning Representations*.

Shen, Y.; Shen, Z.; Wang, M.; Qin, J.; Torr, P.; and Shao, L. 2021. You never cluster alone. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 34, 27734–27746.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 33, 596–608.

Tsai, T. W.; Li, C.; and Zhu, J. 2020. Mice: Mixture of contrastive experts for unsupervised image clustering. In *Proceedings of the International Conference on Learning Representations*.

Van Gansbeke, W.; Vandenhende, S.; Georgoulis, S.; Proesmans, M.; and Van Gool, L. 2020. Scan: Learning to classify images without labels. In *Proceedings of the European Conference on Computer Vision*, 268–285.

Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.-A.; and Bottou, L. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(12).

Wang, X.; Wu, J.; Lin, Z.; Zhang, F.; Zhang, D.; and Nie, L. 2025. Video DataFlywheel: Resolving the Impossible Data Trinity in Video-Language Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ward Jr, J. H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301): 236–244.

Wu, J.; Li, Z.; Sun, W.; Yin, J.; Nie, L.; and Lin, Z. 2025. ClusMatch: Improving Deep Clustering by Unified Positive and Negative Pseudo-label Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wu, J.; Long, K.; Wang, F.; Qian, C.; Li, C.; Lin, Z.; and Zha, H. 2019. Deep comprehensive correlation mining for image clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, 8150–8159.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *Proceedings of the International Conference on Machine Learning*, 478–487.

Xie, X.; Wu, J.; Liu, G.; and Lin, Z. 2024. Sscnet: learning-based subspace clustering. *Visual Intelligence*, 2(1): 11.

Zeiler, M. D.; Krishnan, D.; Taylor, G. W.; and Fergus, R. 2010. Deconvolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2528–2535.

Zelnik-Manor, L.; and Perona, P. 2004. Self-tuning spectral clustering. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 17.

Zhang, C.; Fu, H.; Hu, Q.; Cao, X.; Xie, Y.; Tao, D.; and Xu, D. 2018. Generalized latent multi-view subspace clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(1): 86–99.

Zhang, C.; Fu, H.; Liu, S.; Liu, G.; and Cao, X. 2015. Low-rank tensor constrained multiview subspace clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, 1582–1590.

Zhang, Z.; Zhang, Y.; Li, S.; Liu, G.; Zeng, D.; Yan, S.; and Wang, M. 2019. Flexible auto-weighted local-coordinate concept factorization: A robust framework for unsupervised clustering. *IEEE Transactions on Knowledge and Data Engineering*, 33(4): 1523–1539.

Zhong, H.; Wu, J.; Chen, C.; Huang, J.; Deng, M.; Nie, L.; Lin, Z.; and Hua, X.-S. 2021. Graph contrastive clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, 9224–9233.

Zhou, X.; and Zhang, N. L. 2022. Deep clustering with features from self-supervised pretraining. *arXiv preprint arXiv:2207.13364*.