Semantic-Augmented Image Clustering via Adaptive Multi-Modal Collaboration

Xiaohan Zhang[1], Chao Zhang[1], Deng Xu[1], Hong YU[2], Chunlin Chen[1] and Huaxiong Li[1]*

[1]Nanjing University
[2]Chongqing University of Post and Telecommunications
{xhzhang, chzhang, dengxu}@smail.nju.edu.cn, yuhong@cqupt.edu.cn, {clchen, huaxiongli}@nju.edu.cn

Abstract

Image clustering is a fundamental task in unsupervised visual learning. While recent self-supervised methods have explored various pretext tasks to generate supervision signals for clustering, they typically depend exclusively on raw images, resulting in insufficient supervision signals that are inherently constrained by limited visual semantics. In this paper, we propose a novel Semantic-Augmented image Clustering (SAC) method, which transcends the inherent limitations of purely visual representations through the integration of external knowledge. Specifically, SAC utilizes Vision-Language pre-trained Models (VLMs) to flexibly generate textual descriptions for each image, providing external semantic cues to supplement the visual information. By integrating both visual and textual information, SAC achieves image clustering through a multi-modal learning framework. To mitigate the negative impact of inaccurate textual information, SAC designs an uncertainty-driven adaptive weighting mechanism that explores both intra-modal and inter-modal neighborhood structures, and incorporates the adaptive weights into intra-modal and inter-modal contrastive learning, which improves the robustness against noisy image-text correspondences. Experiments on several popular datasets demonstrate the superiority of SAC compared to state-of-the-art methods.

Introduction

As a longstanding task in unsupervised visual learning, image clustering aims to partition images into groups in the absence of ground-truth labels. Existing image clustering methods can generally be categorized into two types: classical methods and deep learning-based methods. Classical approaches typically rely on prior knowledge or assumptions, such as hand-crafted feature descriptors (Lowe 2004), manifolds (Belkin and Niyogi 2001; Roweis and Saul 2000), sparse (Elhamifar and Vidal 2013; Peng, Zhang, and Yi 2013) and low-rank structures (Liu, Lin, and Yu 2010; Nie et al. 2016; Zhang et al. 2022, 2025), to learn latent representations for clustering. Despite the impressive performance of these methods, they usually struggle to maintain effectiveness when confronted with complex and high-dimensional data. Thanks to the great success of deep learning, deep image clustering methods adopt deep neural networks to extract high-level informative representation for facilitating the downstream clustering tasks (Xie, Girshick, and Farhadi 2016; Caron et al. 2018; Li et al. 2021).

In recent years, self-supervised learning paradigms have been widely adopted in deep clustering, which design various pretext tasks to provide supervision signals for guiding the learning process (Noroozi and Favaro 2016; Gidaris, Singh, and Komodakis 2018; Chen et al. 2020; He et al. 2020, 2022). Among them, contrastive learning has demonstrated promising results in both representation learning and downstream tasks. Its core idea is to maximize the similarities between positive pairs while simultaneously minimizing those between negative pairs. The positive and negative correlations are typically determined by sample similarity or data augmentation, which serve as supervision signals. Despite their remarkable success, most existing methods generate supervision signals in an internal manner, while those signals only rely on given images and are inherently constrained by the limited visual information. Therefore, some recent works attempt to leverage external knowledge, e.g., text, to enhance the image clustering performance (El Banani, Desai, and Johnson 2023; Cai et al. 2023; Li et al. 2024; Qiu et al. 2024). Representatively, Cai et al. (Cai et al. 2023) utilized WordNet (Miller 1995) to assign each image some textual semantic tags, enabling the integration of external linguistic knowledge to guide visual representation learning in a multi-modal manner. Li et al. (Li et al. 2024) and Qiu et al. (Qiu et al. 2024) also adopted WordNet to construct a textual space and performed cross-modal contrastive learning for representation alignment. Given the richer semantics embedded in the textual modality compared to the visual modality, these approaches demonstrate substantial improvements and point to a promising direction for image clustering.

Clearly, the performance of external semantics-guided image clustering hinges on two key aspects: (1) how to generate high-quality external semantics, and (2) how to effectively integrate internal and external semantics for clustering. Although WordNet provides a comprehensive lexical ontology, it contains a vast number of concepts that are irrelevant to the given data. Since the assigned semantic tags heavily rely on the pre-defined semantic vocabulary (Cai et al. 2023; Li et al. 2024; Qiu et al. 2024), this overabundance may degrade the quality of external semantics, even when filtering strategies are applied. Moreover, it is inevitable that some selected

*Corresponding author.

words are incorrect, while the resulting noisy image-text correspondences are often overlooked previously.

In this paper, we propose a new image clustering method with Semantic Augmentation and Adaptive Collaboration, SAC for short. SAC leverages VLMs to flexibly generate textual descriptions for each image, and subsequently integrates both visual and textual modalities to improve image clustering performance through multi-modal collaboration. Considering that the noisy image-text correspondences are caused by wrong or inaccurate textual descriptions, SAC designs an uncertainty-driven adaptive weighting mechanism by exploring the intra- and inter-modal neighborhood structures, and incorporates it into robust contrastive learning. The main contributions of this work are summarized as follows.

- We propose a novel deep image clustering method called SAC, which effectively leverages external knowledge from VLMs and enhances clustering performance through multi-modal collaboration.
- To enhance robustness against noisy image-text correspondences, we mine both intra-modal and inter-modal neighborhood structures for adaptive weighting, and incorporate them into a robust contrastive learning framework.
- Extensive experiments on several popular datasets are conducted to illustrate the superiority of our method over various state-of-the-art image clustering methods.

Related Work

Deep Image Clustering

Early deep clustering methods combine deep representation learning with shallow clustering techniques like $k$means, subspace clustering, or spectral clustering. These methods usually require post-processing to finalize clustering. Recent research has shifted to end-to-end frameworks that directly predict the cluster assignments of images via deep neural networks. One of the major advances in image clustering has been achieved by contrastive learning (Li et al. 2021; Shen et al. 2021; Zhong et al. 2021; Huang et al. 2023), which encourages representations to be similar for positive pairs while pushing apart the representations of negative pairs. Positive samples are often composed of augmented views of the same image or its neighborhood in the feature space. Despite their great success, these methods rely solely on internal information, e.g., input images, whose performance may be limited by insufficient visual semantics. More recently, researchers have begun to shift their attention from internal information to external guidance (Cai et al. 2023; Li et al. 2024; Liu et al. 2024; Qiu et al. 2024). For example, TAC (Li et al. 2024) and SIC (Cai et al. 2023) propose selecting a word caption from WordNet for each image, and then performing image–text cross-modal alignment, which improves clustering performance compared to using images alone. However, they overlook the potential noise in the selected words. Although MCA (Qiu et al. 2024) performs multi-level alignment to alleviate the negative influence of noisy words, accurately estimating the reliability of each image–text pair and fully exploiting the potential of multi-modal collaboration remain open challenges for future research. Moreover, these previous works rely on a fixed WordNet vocabulary, which restricts them from producing more semantically accurate and fine-grained descriptions.

Vision-Language Pre-training Models

The core objective of VLMs is to effectively capture and model interactions between visual and textual modalities. According to the network structure, VLMs are generally categorized into two types: single-stream and dual-stream architectures. The former fuses image and text representations early, enabling deep cross-modal interaction, while the latter processes each modality independently before alignment, offering greater flexibility and efficiency. As a representative dual-stream VLM, CLIP (Radford et al. 2021) employs a ViT-based image encoder and a Transformer-based text encoder, and introduces a contrastive learning framework that aligns image and text embeddings in a shared latent space. Building upon this foundation, BLIP (Li et al. 2022) introduces a unified framework for vision-language understanding and generation, based on a flexible encoder-decoder architecture that supports both discriminative and generative tasks. Subsequently, BLIP-2 (Li et al. 2023) introduces a parameter-efficient design that decouples vision encoding and language modeling. Its lightweight architecture and high efficiency have made it a foundation for many subsequent multi-modal models (Liu et al. 2023). VLMs have demonstrated remarkable versatility across a broad range of tasks, including discriminative tasks such as zero-shot image classification and image-text retrieval, as well as generative tasks like image captioning and visual question answering. Inspired by the powerful generative capacity, we leverage VLMs to generate rich textual information for enhancing the image clustering performance.

Method

In this section, we elaborate on the proposed SAC method, as illustrated in Fig. 1.

Semantic Augmentation Based Clustering

Compared to visual semantics, textual semantics are inherently more effective for discriminative tasks such as classification and clustering. For example, while the words "cat" and "dog" are semantically distinct in textual form, their visual representations may share similar features, making visual discrimination more ambiguous, especially in unsupervised settings. Therefore, our method generates textual counterparts for semantic augmentation. Given an image dataset with n samples, denoted as $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$, we leverage the powerful capabilities of VLMs to bridge the gap between images and text. For each image x_i, we generate a corresponding textual description by a VLM such as BLIP-2 (Li et al. 2023), denoted as

$$y_i = \mathrm{VLM}(x_i). \tag{1}$$

Then, for the image and text modalities, we employ pre-trained encoders to extract the visual embedding $v_i = e_V(x_i)$ and textual embedding $t_i = e_T(y_i)$, where $e_V(\cdot)$ and $e_T(\cdot)$ represent the image encoder and text encoder, respectively.

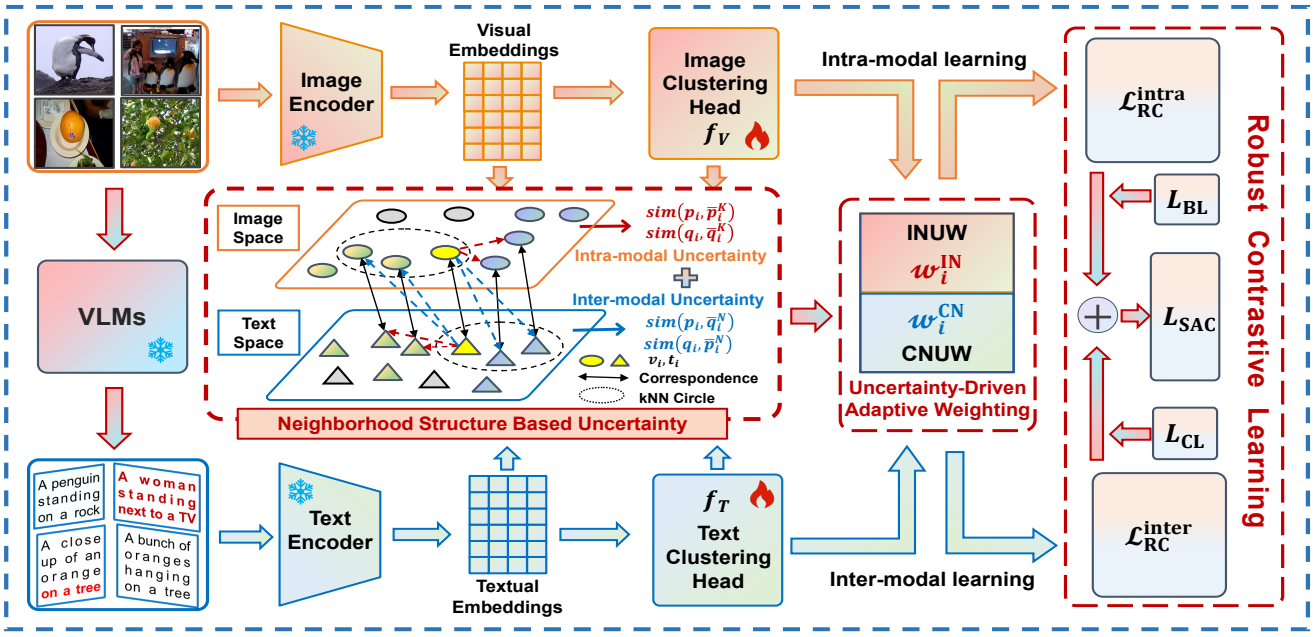


Figure 1: The framework of the proposed SAC. SAC leverages VLMs to generate semantic guidance. To mitigate the negative impact of inaccurate textual information, SAC introduces a contrastive learning scheme with adaptive collaboration.

Both e_V and e_T are kept frozen during the training process. Let c be the number of clusters, f_V: $v \to p \in \mathbb{R}^c$ and f_T: $t \to q \in \mathbb{R}^c$ represent the image and text clustering head, respectively, which map the visual and textual embeddings to soft cluster assignment probability vectors, i.e.,

$$\begin{aligned} p_i &= f_V(v_i; \theta_V) = f_V(e_V(x_i); \theta_V), \\ q_i &= f_T(t_i; \theta_T) = f_T(e_T(y_i); \theta_T), \end{aligned} \tag{2}$$

where θ_V and θ_T are the network parameters.

By well leveraging the constructed image-text pairs, the clustering performance is expected to improve through multi-modal collaboration. A key challenge lies in mitigating the negative impact of noisy image-text correspondences, which arise from wrong or inaccurate text generation. For example, as shown in Fig. 1, VLM produces an inaccurate description "a woman standing next to a TV", which ignores the central subject, penguin.

Contrastive Learning with Adaptive Collaboration

After augmenting the textual view, a natural approach is to apply contrastive learning to the image-text pairs to enhance the instance discrimination. However, the noisy image-text correspondences hinder the effectiveness of contrastive learning. To address this challenge, we design an uncertainty-driven adaptive weighting mechanism and integrate it with both intra-modal and inter-modal contrastive learning for robust clustering.

Uncertainty-Driven Adaptive Weighting To enhance robustness against noisy correspondences, we amplify the contribution of reliable samples while suppressing the impact of noisy ones. Inspired by recent structure-aware advances (Jiang et al. 2025; Zhang et al. 2024), we explore the multi-modal neighborhood structure to measure the uncertainty between each image and its corresponding text.

For the image-text pair (x_i, y_i), we first identify the k nearest neighbors in the textual space based on similarity $\mathrm{sim}(t_i, t_j)$ $(i \neq j)$. The indices of these nearest neighbors in the textual space are denoted as $\mathcal{N}_i^T$.

When an image-text pair (x_i, y_i) is correctly aligned with consistent semantics, its nearest neighbors in the textual space are also expected to correspond to visually similar images, which means that the cluster assignments $\{p_j | j \in \mathcal{N}_i^T\}$ should also be similar to p_i. Thus, we measure the intra-modal similarity between p_i and $\{p_j | j \in \mathcal{N}_i^T\}$, i.e.,

$$w_i^{\mathrm{i2i}} = \mathrm{sgm}\left(\frac{\mathrm{sim}\left(p_i, \bar{p}_i^K\right)}{a}\right), \tag{3}$$

where $\bar{p}_i^K = \frac{1}{k}\sum_{j \in \mathcal{N}_i^T} p_j$ is the average cluster assignment of image modality, $a > 0$ is a scale factor, and sgm is the sigmoid function. Intuitively, if p_i is similar to $\{p_j | j \in \mathcal{N}_i^T\}$, x_i and y_i have more similar neighborhood structures, indicating them more likely to be correctly aligned, and consequently, w_i^{i2i} will be larger. Eq. (3) captures the similarity within image modality. As a complementary measure, we also compute $w_i^{t2t} = \mathrm{sgm}\left(\mathrm{sim}\left(q_i, \bar{q}_i^K\right)/a\right)$ in the text modality, where $\bar{q}_i^K = \frac{1}{k}\sum_{j \in \mathcal{N}_i^V} q_j$ is the average cluster assignment of text modality, and the neighbor set $\{q_j | j \in \mathcal{N}_i^V\}$ is constructed analogously to $\{p_j | j \in \mathcal{N}_i^T\}$. Then, the Intra-modal Neighborhood Uncertainty Weight (INUW) is defined as

$$w_i^{\mathrm{IN}} = \alpha \cdot w_i^{\mathrm{i2i}} + (1 - \alpha) \cdot w_i^{\mathrm{t2t}}, \tag{4}$$

where $\alpha \in (0, 1)$ is a balance parameter. A larger w_i^{IN} indicates greater intra-modal neighborhood consistency between

two modalities.

The weight w_i^{IN} characterizes the consistency of intra-modal neighborhood structures, and we further explore the consistency of inter-modal neighborhood structures. If the image-text pair (x_i, y_i) is correctly aligned, the cluster assignment of the image modality p_i should be similar to that of its textual neighbors $\{q_j | j \in \mathcal{N}_i^T\}$; conversely, the cluster assignment of the text modality q_i should be similar to that of its visual neighbors $\{p_j | j \in \mathcal{N}_i^V\}$. Thus, we can directly measure the cross-modal neighborhood similarity by

$$w_i^{\text{i2t}} = \text{sgm}\left(\frac{\text{sim}\left(p_i, \bar{q}_i^N\right)}{a}\right), w_i^{\text{t2i}} = \text{sgm}\left(\frac{\text{sim}\left(q_i, \bar{p}_i^N\right)}{a}\right), \tag{5}$$

where $\bar{q}_i^N = \frac{1}{k}\sum_{j\in\mathcal{N}_i^T} q_j$ and $\bar{p}_i^N = \frac{1}{k}\sum_{j\in\mathcal{N}_i^V} p_j$. Then, the Cross-modal Neighborhood Uncertainty Weight (CNUW) is defined as

$$w_i^{\text{CN}} = \alpha \cdot w_i^{\text{i2t}} + (1-\alpha)\cdot w_i^{\text{t2i}}. \tag{6}$$

A larger w_i^{CN} indicates the higher cross-modal neighborhood consistency.

Robust Contrastive Learning Based on the INUW and CNUW, we incorporate them into robust intra-modal and cross-modal contrastive learning to improve the model's robustness against noisy correspondences. The robust contrastive loss is composed of

$$\mathcal{L}_{\text{RC}} = \mathcal{L}_{\text{RC}}^{\text{intra}} + \mathcal{L}_{\text{RC}}^{\text{inter}}, \tag{7}$$

where $\mathcal{L}_{\text{RC}}^{\text{intra}}$ and $\mathcal{L}_{\text{RC}}^{\text{inter}}$ denote the intra-modal contrastive learning loss and inter-modal contrastive learning loss, respectively. $\mathcal{L}_{\text{RC}}^{\text{intra}}$ is defined as

$$\mathcal{L}_{\text{RC}}^{\text{intra}} = -\sum_{i}^{n} w_i^{\text{IN}} \cdot \log \frac{e^{\text{sim}\left(p_i, \bar{p}_i^K\right)/\tau}}{e^{\text{sim}\left(p_i, \bar{p}_i^K\right)/\tau} + \sum_{j\neq i} e^{\text{sim}\left(p_i, \bar{p}_j^K\right)/\tau}}. \tag{8}$$

$\mathcal{L}_{\text{RC}}^{\text{inter}}$ is a bidirectional loss defined as

$$\mathcal{L}_{\text{RC}}^{\text{inter}} = \sum_{i=1}^{n} \mathcal{L}_i^{i\to t} + \sum_{i=1}^{n} \mathcal{L}_i^{t\to i}, \tag{9}$$

$$\mathcal{L}_i^{i\to t} = -w_i^{\text{CN}} \cdot \log \frac{e^{\text{sim}\left(p_i, \bar{q}_i^N\right)/\tau}}{e^{\text{sim}\left(p_i, \bar{q}_i^N\right)/\tau} + \sum_{j\neq i} e^{\text{sim}\left(p_i, \bar{q}_j^N\right)/\tau}}, \tag{10}$$

$$\mathcal{L}_i^{t\to i} = -w_i^{\text{CN}} \cdot \log \frac{e^{\text{sim}\left(q_i, \bar{p}_i^N\right)/\tau}}{e^{\text{sim}\left(q_i, \bar{p}_i^N\right)/\tau} + \sum_{j\neq i} e^{\text{sim}\left(q_i, \bar{p}_j^N\right)/\tau}}, \tag{11}$$

where τ is a temperature parameter. The robust contrastive loss distills the neighborhood information both within and across modalities with adaptive weights for discriminative representation learning.

Additionally, we introduce two regularization terms to stabilize the training process. Denote $P = [p_1; \cdots; p_n]$ and $Q = [q_1; \cdots; q_n]$, and let $\hat{p}_i \in \mathbb{R}^n$ and $\hat{q}_i \in \mathbb{R}^n$ be the i-th column of P and Q, respectively. We first introduce the bi-directional alignment regularization:

$$\mathcal{L}_{\text{CL}} = \frac{1}{c}\sum_{i=1}^{c} \text{CE}\left(\hat{p}_i, \hat{q}_i\right) + \frac{1}{n}\sum_{i=1}^{n} \text{CE}\left(p_i, q_i\right), \tag{12}$$

Dataset	Image Size	# Training	# Test	# Classes
STL-10	96×96	5,000	8,000	10
CIFAR-10	32×32	50,000	10,000	10
CIFAR-20	32×32	50,000	10,000	20
ImageNet-10	224×224	13,000	500	10
ImageNet-Dogs	64×64	19,500	750	15

Table 1: A summary of datasets used for evaluation.

where CE is the cross-entropy loss. Second, to prevent all samples from collapsing into only a few clusters, we adopt the following balance loss:

$$\mathcal{L}_{\text{BL}} = -\sum_{i=1}^{c} \left(\bar{p}_i \log \bar{p}_i + \bar{q}_i \log \bar{q}_i\right), \tag{13}$$

where $\bar{p}_i = \frac{1}{n}\sum_i p_i$, and $\bar{q}_i = \frac{1}{n}\sum_i q_i$.

To this end, the overall objective function of SAC can be formulated as:

$$\mathcal{L}_{\text{SAC}} = \mathcal{L}_{\text{RC}} + \lambda_a \cdot \mathcal{L}_{\text{CL}} - \lambda_b \cdot L_{\text{BL}}, \tag{14}$$

where λ_a and λ_b are two trade-off parameters.

Experiments

Experimental Setup

Datasets To assess the effectiveness of our SAC method, we conduct experiments on five standard image clustering benchmark datasets: STL-10 (Coates, Ng, and Lee 2011), CIFAR-10/20 (Krizhevsky and Hinton 2009), ImageNet-10, and ImageNet-Dogs (Chang et al. 2017). The general statistics of these datasets are presented in Table 1.

Baselines We compare our SAC with 17 state-of-the-art clustering methods on the five datasets, including JULE (Yang, Parikh, and Batra 2016), DEC (Xie, Girshick, and Farhadi 2016), DAC (Chang et al. 2017), DCCM (Wu et al. 2019), IIC (Ji, Henriques, and Vedaldi 2019), PICA (Huang, Gong, and Zhu 2020), CC (Li et al. 2021), IDFD (Tao, Takagi, and Nakata 2020), SCAN (Van et al. 2020), MiCE (Tsai, Li, and Zhu 2020), GCC (Zhong et al. 2021), NNM (Dang et al. 2021), TCC (Shen et al. 2021), SPICE (Niu, Shan, and Wang 2022), SIC (Cai et al. 2023), TAC (Li et al. 2024) and MCA (Qiu et al. 2024). It is worth noting that SIC, TAC, and MCA are recent image clustering approaches that incorporate external text information, whereas the other methods rely solely on visual information.

Evaluation Metrics To comprehensively assess clustering quality, we employ three popular metrics: Normalized Mutual Information (NMI), Clustering Accuracy (ACC) and Adjusted Rand Index (ARI) as evaluation criteria. Higher values on all metrics indicate better performance.

Implementation Details Our framework employs a CLIP ViT-B/32 backbone (Dosovitskiy et al. 2020) to extract 512-dimensional visual embeddings $v_i \in \mathbb{R}^{512}$ of input images. Textual descriptions are generated by BLIP-2 (Li et al. 2023) through its image-to-text decoder, then encoded

Dataset	STL-10			CIFAR-10			CIFAR-20			ImageNet-10			ImageNet-Dogs			AVG
Metrics	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	
JULE (Yang, Parikh, and Batra 2016)	18.2	27.7	16.4	19.2	27.2	13.8	10.3	13.7	3.3	17.5	30.0	13.8	5.4	13.8	2.8	15.5
DEC (Xie, Girshick, and Farhadi 2016)	27.6	35.9	18.6	25.7	30.1	16.1	13.6	18.5	5.0	28.2	38.1	20.3	12.2	19.5	7.9	21.2
DAC (Chang et al. 2017)	36.6	47.0	25.7	39.6	52.2	30.6	18.5	23.8	8.8	39.4	52.7	30.2	21.9	27.5	11.1	31.0
DCCM (Wu et al. 2019)	37.6	48.2	26.2	49.6	62.3	40.8	28.5	32.7	17.3	60.8	71.0	55.5	32.1	38.3	18.2	41.3
IIC (Ji, Henriques, and Vedaldi 2019)	49.6	59.6	39.7	51.3	61.7	41.1	22.5	25.7	11.7	–	–	–	–	–	–	–
PICA (Huang, Gong, and Zhu 2020)	61.1	71.3	53.1	59.1	69.6	51.2	31.0	33.7	17.1	80.2	87.0	76.1	35.2	35.3	20.1	52.1
CC (Li et al. 2021)	76.4	85.0	72.6	70.5	79.0	63.7	43.1	42.9	26.6	85.9	89.3	82.2	44.5	42.9	27.4	62.1
IDFD (Tao, Takagi, and Nakata 2020)	64.3	75.6	57.5	71.1	81.5	66.3	42.6	42.5	26.4	89.8	95.4	90.1	54.6	59.1	41.3	63.9
SCAN (Van et al. 2020)	69.8	80.9	64.6	79.7	88.3	77.2	48.6	50.7	33.3	–	–	–	61.2	59.3	45.7	–
MiCE (Tsai, Li, and Zhu 2020)	63.5	75.2	57.5	73.7	83.5	69.8	43.6	44.0	28.0	–	–	–	42.3	43.9	28.6	–
GCC (Zhong et al. 2021)	68.4	78.8	63.1	76.4	85.6	72.8	47.2	47.2	30.5	84.2	90.1	82.2	49.0	52.6	36.2	64.3
NNM (Dang et al. 2021)	66.3	76.8	59.6	73.7	83.7	69.4	48.0	45.9	30.2	–	–	–	60.4	58.6	44.9	–
TCC (Shen et al. 2021)	73.2	81.4	68.9	79.0	90.6	73.3	47.9	49.1	31.2	84.8	89.7	82.5	55.4	59.5	41.7	67.2
SPICE (Niu, Shan, and Wang 2022)	81.7	90.8	81.2	73.4	83.8	70.5	44.8	46.8	29.4	82.8	92.1	83.6	57.2	64.6	47.9	68.7
SIC (Cai et al. 2023)	95.3	98.1	95.9	84.7	92.6	84.4	59.3	58.3	43.9	97.0	98.2	96.1	69.0	69.7	55.8	79.9
MCA (Qiu et al. 2024)	95.5	98.2	96.0	85.0	92.8	84.9	60.6	61.2	45.5	-	-	-	75.1	77.9	64.3	-
TAC (Li et al. 2024)	95.5	98.2	96.1	83.3	91.9	83.1	61.1	60.7	44.8	98.5	99.2	98.3	80.6	83.0	72.2	83.2
SAC-I (k-means)	88.7	97.5	82.8	70.4	82.7	59.1	46.0	39.2	28.7	92.2	98.8	88.4	39.4	44.1	25.5	70.0
SAC-T (k-means)	88.9	96.8	80.4	86.3	**95.6**	81.3	58.6	55.1	41.3	87.5	96.6	84.3	59.3	59.4	40.1	74.1
SAC	**96.2**	**98.5**	**96.7**	**88.5**	94.9	**89.1**	**66.1**	**67.4**	**51.9**	**99.1**	**99.6**	**99.1**	**83.7**	**86.5**	**75.9**	**86.2**

Table 2: Clustering performance (%) of different methods on five image datasets. The best and second best results are denoted in **bold** and underline, respectively.

Noise	STL-10			CIFAR-10			CIFAR-20			ImageNet-10			ImageNet-Dogs			AVG
	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	
0%	**96.2**	**98.5**	**96.7**	**88.5**	**94.9**	**89.1**	**66.1**	**67.4**	**51.9**	**99.1**	**99.6**	**99.1**	**83.7**	**86.5**	**75.9**	**86.2**
30%	**96.2**	**98.5**	**96.7**	88.0	94.6	88.5	66.0	64.5	50.2	**99.1**	**99.6**	**99.1**	76.9	76.5	65.7	84.0
60%	95.9	98.4	96.5	86.9	94.0	87.4	63.1	62.2	47.3	98.7	99.4	98.6	75.1	76.1	63.5	82.9

Table 3: Clustering performance of SAC under different noise rates.

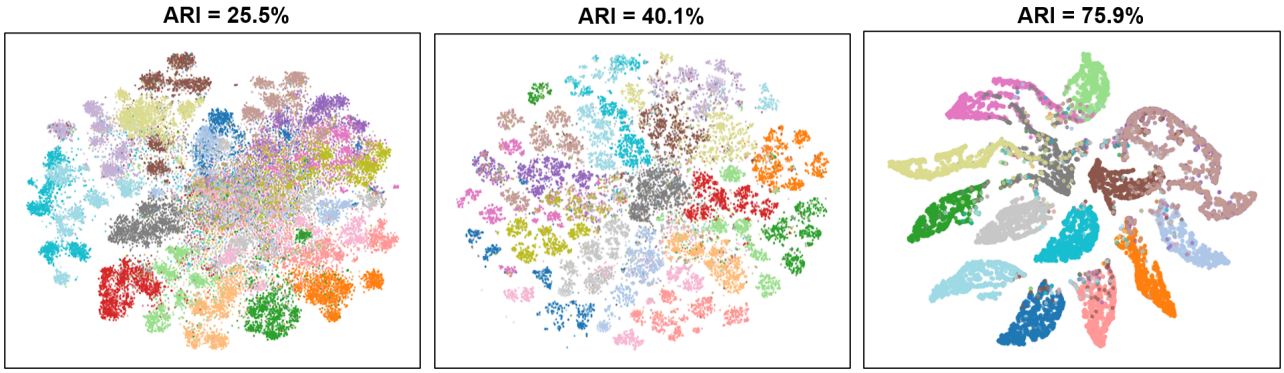
into 768-dimensional semantic vectors $t_i \in \mathbb{R}^{768}$ using SBERT (Reimers and Gurevych 2019) to get pure textual representations. Both image and text encoders are frozen during the training process. The only trainable component is the image/text clustering head implemented as a fully-connected layer $f_c : \mathbb{R}^{512/768} \to \mathbb{R}^c$, where c denotes the cluster number.

We optimize the model using Adam optimizer ($\eta = 10^{-3}$, batch size = 512) with early-stopping. The hyper-parameters are set to $\tau = 1.1$, $a = 1$, $\alpha = 0.9$, $\lambda_a = 0.6$ and $\lambda_b = 4$ across all datasets. All experiments are executed with an NVIDIA A100 GPU.

Main Results

Comparisons with State-of-the-arts Table 2 lists the clustering results of different methods on all datasets w.r.t. three metrics. From the results, we can draw the following observations: (1) Compared to methods that rely solely on visual information (from JULE to SPICE), our approach achieves significant improvements in clustering performance. For example, on the STL-10 dataset, SPICE achieves an ACC of 90.8%, whereas SAC reaches 98.5%. The results demonstrate the effectiveness of incorporating external textual information for enhancing visual representation learning. (2) Compared to SIC, MCA, and TAC that also leverage text information, SAC consistently outperforms them, especially on CIFAR-10, CIFAR-20, and ImageNet-Dogs datasets. The main reason is that our method leverages VLMs to improve the quality of texts and designs a robust contrastive learning framework for clustering. (3) To further investigate the effectiveness of multi-modal collaboration, we design SAC-I and SAC-T, which perform $k$means on the extracted visual embeddings and textual embeddings, respectively. SAC-T significantly outperforms SAC-I in most cases such as CIFAR-10, CIFAR-20 and ImageNet-Dogs datasets, demonstrating that the generated textual descriptions capture richer semantic information. SAC outperforms the two variants, validating the effectiveness of multi-modal collaboration.

Robustness against Noisy Correspondences To evaluate the robustness of SAC against noisy image-text correspondences, we conduct experiments under three conditions: no additional noise, 30% artificial noise, and 60% artificial noise. The artificial noise is introduced by randomly deleting words within a sentence and mixing sentences generated from two different images at a pre-defined rate. The clustering results are presented in Table 3. SAC maintains relative stable perfor-



(a) Visual embeddings (b) Textual embeddings (c) Image cluster assignments

Figure 2: t-SNE visualization of different features on the ImageNet-Dogs training set.

Loss Terms			CIFAR-20			ImageNet-Dogs		
$\mathcal{L}_{\text{RC}}$	$\mathcal{L}_{\text{BL}}$	$\mathcal{L}_{\text{CL}}$	NMI	ACC	ARI	NMI	ACC	ARI
✓	-	-	56.6	28.7	30.0	15.9	7.8	5.3
-	✓	-	12.5	5.0	1.5	64.8	9.2	40.0
-	-	✓	58.9	57.0	39.9	68.2	66.4	51.0
-	✓	✓	57.2	56.2	39.4	56.7	48.7	34.1
✓	-	✓	57.2	56.3	39.4	81.1	72.8	62.8
✓	✓	-	**66.1**	66.2	61.8	82.0	84.4	72.9
✓	✓	✓	**66.1**	**67.4**	**51.9**	**83.7**	**86.5**	**75.9**

Table 4: Performance of SAC with different loss terms.

Weights		CIFAR-20			ImageNet-Dogs		
w^{IN}	w^{CN}	NMI	ACC	ARI	NMI	ACC	ARI
-	-	65.2	65.7	49.4	82.9	84.9	73.7
✓	-	65.4	66.1	49.8	82.9	86.1	74.8
-	✓	65.2	66.8	49.9	82.9	85.9	74.6
✓	✓	**66.1**	**67.4**	**51.9**	**83.7**	**86.5**	**75.9**

Table 5: Performance of SAC with different weights.

Components		CIFAR-20			ImageNet-Dogs		
$\mathcal{L}_{\text{RC}}^{\text{intra}}$	$\mathcal{L}_{\text{RC}}^{\text{inter}}$	NMI	ACC	ARI	NMI	ACC	ARI
-	✓	64.9	64.1	50.0	80.0	78.1	68.6
✓	-	65.3	66.4	51.2	82.9	85.4	74.3
✓	✓	**66.1**	**67.4**	**51.9**	**83.7**	**86.5**	**75.9**

Table 6: Performance of SAC with different contrastive loss.

mance across varying levels of additional noise, especially on STL-10 and ImageNet-10 datasets. The possible reason is that the adopted VLM is capable of generating accurate textual descriptions, and our robust contrastive learning framework further enhances resistance to noisy correspondences.

Visualization To illustrate the effectiveness of text-guided mechanisms in refining feature distributions for image clustering, Fig. 2 presents a comparative visualization of final cluster assignments alongside the visual and textual embeddings extracted by their respective encoders on the ImageNet-Dogs dataset. The visual embeddings show relatively poor separation among clusters, while the textual embeddings exhibit improved cluster separation, reflecting the richer semantic information. By leveraging both modalities through adaptive collaboration, the final learned cluster assignments achieve the most distinct and compact clustering, demonstrating the effectiveness of the text-guided multi-modal collaboration in improving image clustering performance.

Ablation Studies

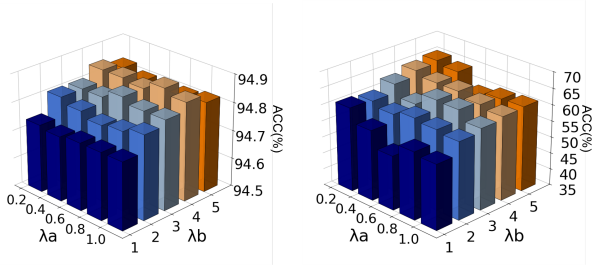
Loss Terms An ablation study on the loss components $\mathcal{L}_{\text{RC}}$, $\mathcal{L}_{\text{CL}}$, and $\mathcal{L}_{\text{BL}}$ is conducted on CIFAR-20 and ImageNet-Dogs datasets to evaluate their effectiveness. In our method, $\mathcal{L}_{\text{RC}}$ distills both intra-modal and inter-modal neighborhood information, $\mathcal{L}_{\text{BL}}$ prevents cluster collapse by promoting balanced assignments, and $\mathcal{L}_{\text{CL}}$ facilitates cross-modal bi-directional alignment at both category and instance levels. As shown in Table 4, each of these loss components contributes to improving the clustering performance.

Adaptive Weights Table 5 presents the clustering results of SAC with different adaptive weights, i.e., w^{IN} and w^{CN} in $\mathcal{L}_{\text{CL}}$, on CIAFR-20 and ImageNet-Dogs datasets. When neither w^{IN} nor w^{CN} is applied, $\mathcal{L}_{\text{RC}}$ reduces to a standard contrastive learning form, yielding the lowest performance. By incorporating both intra-modal and inter-modal neighborhood uncertainty weights, our method guides contrastive learning to focus more effectively on well-aligned sample pairs, thereby achieving the best performance.

Intra-modal and Inter-modal Contrastive learning Our robust contrastive learning includes intra-modal and inter-modal learning. To evaluate their effectiveness, we compare the performance of SAC with different robust contrastive loss on CIFAR-20 and ImageNet-Dogs, and report the clustering results in Table 7. The results show that combining intra-modal and inter-modal contrastive learning yields the best performance, confirming that intra-modal and inter-modal learning provide complementary signals, structural consistency

Generator	CIFAR-10			CIFAR-20		
	NMI	ACC	ARI	NMI	ACC	ARI
WordNet	85.4	93.2	85.8	59.8	55.7	42.4
BLIP	88.1	94.6	88.7	58.6	55.8	42.6
BLIP-2	**88.5**	**94.9**	**89.1**	**66.1**	**67.4**	**51.9**

Table 7: Performance of SAC with different text generators.



(a) CIFAR-10 (b) CIFAR-20

Figure 3: Sensitivity analysis of λ_a and λ_b.

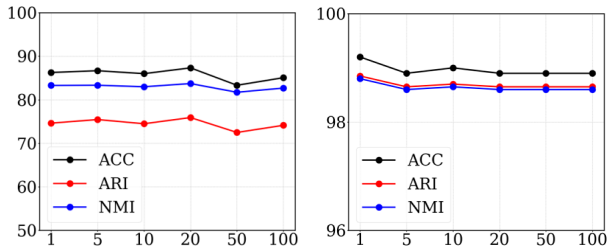
within modalities, and semantic alignment across modalities.

Text Generator Our SAC adopts BLIP-2 as a text generator to generate the textual descriptions for each image. To evaluate the impact of different text generators, we conduct experiments comparing WordNet, BLIP, and BLIP-2. For WordNet, we adopt the same strategy as in TAC (Li et al. 2024) to select descriptive words. For BLIP and BLIP-2, the textual sentences are generated using their respective image-to-text decoders. As shown in Table 7, SAC using VLMs (BLIP and BLIP-2) generally achieves better performance compared to using WordNet. The main reason is that WordNet relies on selecting semantically similar words from a fixed vocabulary, which constrains its flexibility and expressive capacity. In contrast, BLIP and BLIP-2 generate image-conditioned descriptions with richer contextual information. Besides, BLIP-2 achieves better results than BLIP, owing to its stronger vision-language understanding capabilities. These results indicate that generating more accurate textual descriptions can further enhance clustering performance.

Parameter Analyses

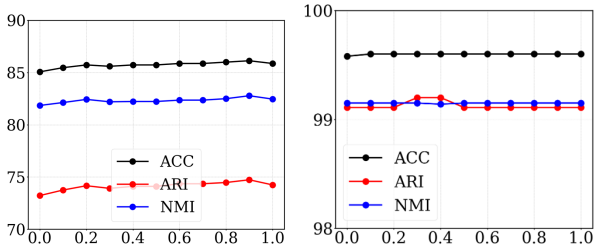
Parameters λ_a and λ_b We investigate the impact of the parameters λ_a and λ_b, where λ_a encourages bi-directional clustering consistency across modalities, and λ_b mitigates cluster collapse while promoting balanced category distributions. Fig. 3 shows the ACC of SAC with different parameters on two datasets. The optimal parameters vary across datasets. In general, SAC can achieve relatively stable and satisfactory performance when λ_a and λ_b are selected from $[0.4, 0.8]$ and $[2, 4]$, respectively. In our experiments, we set the default $\lambda_a = 0.6$ and $\lambda_b = 4$ for simplicity.

Number of Neighbors k SAC distills the neighborhood information both within and across modalities with adaptive contrastive weights. Here we evaluate the impact of the number of nearest neighbors k. As shown in Fig. 4, the clustering performance of SAC is not very sensitive to k. Although the optimal k value varies across datasets, we set the default $k = 20$ on all datasets for simplicity, which also achieves stable and satisfactory results.



(a) ImageNet-Dogs (b) ImageNet-10

Figure 4: Sensitivity analysis of Neighbors' number k.



(a) ImageNet-Dogs (b) ImageNet-10

Figure 5: Sensitivity analysis of α.

Balance Factor α The balance factor α in INUW and CNUW controls the relative importance of the neighbor information within two modalities. A larger α value increases reliance on the textual neighborhoods to determine the uncertainty weight. As shown in Fig. 5, a large α value leads to improved performance on ImageNet-Dogs, while its impact is less pronounced on ImageNet-10. We set the default $\alpha = 0.9$ across all datasets.

Conclusion

This paper proposes a novel SAC method that tackles the image clustering problem in a multi-modal fashion by introducing external knowledge from VLMs. SAC utilizes VLMs to generate textual descriptions for each image, thereby enhancing semantic information and offering complementary perspectives. To mitigate the negative impact of inaccurate textual information, SAC designs a robust contrastive learning framework with an uncertainty-driven adaptive weighting mechanism. It explores the neighborhood structure within and across modalities, and distills the neighborhood information with adaptive weights. Experimental results on datasets validate its effectiveness and superiority of the state-of-the-art approaches. .

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grants Nos. 62576161, 62176116, and 62276136.

References

Belkin, M.; and Niyogi, P. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. *NIPS*, 14: 585–591.

Cai, S.; Qiu, L.; Chen, X.; Zhang, Q.; and Chen, L. 2023. Semantic-enhanced image clustering. In *AAAI*, volume 37, 6869–6878.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep clustering for unsupervised learning of visual features. In *ECCV*, 132–149.

Chang, J.; Wang, L.; Meng, G.; Xiang, S.; and Pan, C. 2017. Deep adaptive image clustering. In *ICCV*, 5879–5887.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607.

Coates, A.; Ng, A.; and Lee, H. 2011. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, 215–223.

Dang, Z.; Deng, C.; Yang, X.; Wei, K.; and Huang, H. 2021. Nearest neighbor matching for deep clustering. In *CVPR*, 13693–13702.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

El Banani, M.; Desai, K.; and Johnson, J. 2023. Learning visual representations via language-guided sampling. In *CVPR*, 19208–19220.

Elhamifar, E.; and Vidal, R. 2013. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE TPAMI*, 35(11): 2765–2781.

Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *CVPR*, 16000–16009.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 9729–9738.

Huang, J.; Gong, S.; and Zhu, X. 2020. Deep semantic clustering by partition confidence maximisation. In *CVPR*.

Huang, Z.; Chen, H.; Wen, Z.; Zhang, C.; Li, H.; Wang, B.; and Chen, C. 2023. Model-aware contrastive learning: Towards escaping the dilemmas. In *ICML*, 13774–13790. PMLR.

Ji, X.; Henriques, J. F.; and Vedaldi, A. 2019. Invariant information clustering for unsupervised image classification and segmentation. In *ICCV*, 9865–9874.

Jiang, B.; Zhang, C.; Wang, Z.; Liang, X.; Zhou, P.; Du, L.; Zhang, Q.; Ding, W.; and Liu, Y. 2025. Scalable fuzzy clustering with collaborative structure learning and preservation. *IEEE Transactions on Fuzzy Systems*, 33(9): 3047–3060.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 19730–19742.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 12888–12900.

Li, Y.; Hu, P.; Liu, Z.; Peng, D.; Zhou, J. T.; and Peng, X. 2021. Contrastive clustering. In *AAAI*, volume 35, 8547–8555.

Li, Y.; Hu, P.; Peng, D.; Lv, J.; Fan, J.; and Peng, X. 2024. Image clustering with external guidance. In *ICML*, 27890–27902.

Liu, G.; Lin, Z.; and Yu, Y. 2010. Robust subspace segmentation by low-rank representation. In *ICML*, 663–670.

Liu, H.; Hu, P.; Zhang, C.; Li, Y.; and Peng, X. 2024. Interactive deep clustering via value mining. *NeurIPS*, 37: 42369–42387.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *NIPS*, 36: 34892–34916.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *IJCV*, 60: 91–110.

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

Nie, F.; Wang, X.; Jordan, M.; and Huang, H. 2016. The constrained laplacian rank algorithm for graph-based clustering. In *AAAI*, volume 30, 1969–1976.

Niu, C.; Shan, H.; and Wang, G. 2022. Spice: semantic pseudo-labeling for image clustering. *IEEE TIP*, 31: 7264–7278.

Noroozi, M.; and Favaro, P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 69–84. Springer.

Peng, X.; Zhang, L.; and Yi, Z. 2013. Scalable sparse subspace clustering. In *CVPR*, 430–437.

Qiu, L.; Zhang, Q.; Chen, X.; and Cai, S. 2024. Multi-level cross-modal alignment for image clustering. In *AAAI*, volume 38, 14695–14703.

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 *ICML*, 8748–8763.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: sentence embeddings using siamese bert-networks. In *EMNLP*, 3982–3992.

Roweis, S. T.; and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500): 2323–2326.

Shen, Y.; Shen, Z.; Wang, M.; Qin, J.; Torr, P.; and Shao, L. 2021. You never cluster alone. *NIPS*, 34: 27734–27746.

Tao, Y.; Takagi, K.; and Nakata, K. 2020. Clustering-friendly Representation Learning via Instance Discrimination and Feature Decorrelation. In *ICLR*.

Tsai, T. W.; Li, C.; and Zhu, J. 2020. Mice: mixture of contrastive experts for unsupervised image clustering. In *ICLR*.

Van, G. W.; Vandenhende, S.; Georgoulis, S.; Proesmans, M.; and Van Gool, L. 2020. Scan: learning to classify images without labels. In *ECCV*, 268–285. Springer.

Wu, J.; Long, K.; Wang, F.; Qian, C.; Li, C.; Lin, Z.; and Zha, H. 2019. Deep comprehensive correlation mining for image clustering. In *ICCV*, 8150–8159.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *ICML*, 478–487.

Yang, J.; Parikh, D.; and Batra, D. 2016. Joint unsupervised learning of deep representations and image clusters. In *CVPR*, 5147–5156.

Zhang, C.; Fang, Y.; Liang, X.; Zhang, H.; Zhou, P.; Wu, X.; Yang, J.; Jiang, B.; and Sheng, W. 2024. Efficient Multi view Unsupervised Feature Selection with Adaptive Structure Learning and Inference. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence.

Zhang, C.; Li, H.; Chen, C.; Jia, X.; and Chen, C. 2022. Low-rank tensor regularized views recovery for incomplete multiview clustering. *IEEE TNNLS*, 35(7): 9312–9324.

Zhang, C.; Wang, Z.; Jia, X.; Li, Z.; Chen, C.; and Li, H. 2025. Multi-view Clustering with Incremental Instances and Views. *IEEE TIP*.

Zhong, H.; Wu, J.; Chen, C.; Huang, J.; Deng, M.; Nie, L.; Lin, Z.; and Hua, X.-S. 2021. Graph contrastive clustering. In *ICCV*, 9224–9233.