Reducing Bias and Variance: Generative Semantic Guidance and Bi-Layer Ensemble for Image Clustering

Feijiang Li[1,2,3*], **Zhenxiong Li**[1], **Jieting Wang**[1,2,3]
Zizheng Jiu[1], **Saixiong Liu**[1], **Liang Du**[1,2,3]
[1]Institute of Big Data Science and Industry, Shanxi University
[2]Key Laboratory of Evolutionary Science Intelligence of Shanxi Province
[3]School of Artificial Intelligence, Shanxi University
feijiangli@163.com, {lizhenxiong, jtwang, jiuzizheng, duliang}@sxu.edu.cn, liu_saixiong@126.com

Abstract

Image clustering aims to partition unlabeled image datasets into distinct groups. A core aspect of this task is constructing and leveraging prior knowledge to guide the clustering process. Recent approaches introduce semantic descriptions as prior information, most of which typically relying on matching-based techniques with predefined vocabularies. However, the limited matching space restricts their adaptability to downstream clustering tasks. Moreover, these methods primarily focus on reducing bias to improve performance, frequently overlooking the importance of variance reduction. To address these limitations, we propose GSEC (Image Clustering based on Generative Semantic Guidance and Bi-Layer Ensemble), a framework designed to reduce bias through generative semantic guidance and mitigate variance via ensemble learning. Our method employs Multimodal Large Language Models to generate semantic descriptions and derive image embeddings via weighted averaging. Additionally, a bi-layer ensemble strategy integrates cross-modal information through BatchEnsemble in the inner layer and aligns outputs via an alignment mechanism in the outer layer. Comparative experiments demonstrate that GSEC outperforms 20 state-of-the-art methods across six benchmark datasets, while further analysis confirms its effectiveness in simultaneously reducing both bias and variance. The code is available at https://github.com/2017LI/GSEC.git. The appendix is available at https://arxiv.org/abs/2605.12961.

1 Introduction

The goal of image clustering is to partition a dataset into multiple disjoint subsets by learning from unlabeled image data. A core aspect of this task is constructing and leveraging prior knowledge to form effective supervisory signals.

Based on the nature of the prior information, image clustering methods can be broadly categorized into several types. Classical clustering approaches rely heavily on assumptions about data distributions, including prototype-based methods [McQueen, 1967], density-based methods [Ester *et al.*, 1996], and hierarchical methods [Dasgupta, 2002]. Deep clustering methods leverage powerful representation learning to extract highly discriminative features as prior information [Xie *et al.*, 2016; Caron *et al.*, 2018]. Advances in self-supervised learning have further shifted clustering paradigms toward supervision signals derived from data augmentation [Li *et al.*, 2021b]. Nevertheless, these approaches rely exclusively on intrinsic information within the data to guide clustering. As a result, the supervisory signals are inherently constrained by the data itself [Liu *et al.*, 2024a], making further performance improvements increasingly difficult in practice.

Inspired by the zero-shot classification paradigm introduced by CLIP [Radford *et al.*, 2021], recent studies have shifted the source of supervision from internal priors to external semantic descriptions. Methods such as SIC [Cai *et al.*, 2023] and TAC [Li *et al.*, 2023b] typically adopt predefined vocabularies (e.g., WordNet) and employ pre-trained vision–language models to embed images and text into a shared representation space. Then, images are matched with the most relevant textual labels to generate semantic descriptions, which serve as external semantic priors to guide the clustering process. Most of these methods construct semantic prior information by matching images with textual libraries. However, due to the limited coverage of predefined vocabularies, forced alignment between images and text may lead to semantic inconsistency, particularly when complex visual semantics surpass the representational capacity of the fixed lexical set. Therefore, these approaches may be constrained by their reliance on matching-based semantic prior construction.

Moreover, the strategy of introducing semantic priors is essentially to improve the clustering performance by reducing the bias. In machine learning theory, as is well known, estimation error arises from the joint effects of bias, variance, and noise. Therefore, explicitly accounting for variance reduction during clustering is expected to further improve performance.

Based on the discussions above, we propose Image Clustering based on Generative Semantic Guidance and Bi-Layer Ensemble, noted as GSEC. This framework jointly improves clustering performance by addressing both bias and variance. To reduce bias, we introduce a generative seman-

tic prior construction strategy that overcomes the limitations of matching-based semantic priors. This strategy leverages Multimodal Large Language Models to generate semantic descriptions and synthesizes corresponding semantic representations for each image via weighted averaging. To reduce variance, we incorporate ensemble learning through a bi-layer ensemble strategy, comprising an inner-layer ensemble and an outer-layer ensemble. The inner layer extracts domain information from both semantic and image modalities and performs internal ensemble via BatchEnsemble to achieve semantically guided image clustering. The outer layer employs an alignment-based ensemble mechanism to align the inner outputs with the outer outputs, with the aligned result serving as the final clustering outcome.

Our contributions are summarized as follows:

- Generative Semantic Guidance: We propose a bias reduction strategy that leverages multimodal large language model to generate adaptive semantic descriptions for images, thereby reducing the inconsistency of the match-based method.
- Bi-Layer Ensemble Mechanism: We design a bi-layer ensemble framework to mitigate variance. The inner ensemble integrates multimodal information via BatchEnsemble, and the outer ensemble aligns inner predictions with external outputs, thereby enhancing robustness and stability.
- Comprehensive Empirical Validation: We conduct experiments on benchmark datasets, demonstrating that GSEC significantly outperforms the reference methods. The bias-variance decomposition analysis confirms that our approach effectively reduces both bias and variance.

2 Related Works

In this section, we briefly review image clustering and ensemble learning.

2.1 Image Clustering

The evolution of image clustering methodologies has progressively incorporated richer forms of prior knowledge to construct more effective supervisory signals [Ren *et al.*, 2024]. In the following, we review these methods according to their primary sources of supervision: internal prior information and external semantic information.

The development of the internal prior information-based clustering algorithms can be broadly divided into three stages. Early classical clustering methods typically rely on low-level visual semantics: K-means [McQueen, 1967] groups data by minimizing distances to cluster centers, while DBSCAN [Ester *et al.*, 1996] identifies arbitrarily shaped clusters based on local density. With the advent of deep learning, deep clustering methods emerged: DEC [Xie *et al.*, 2016] learns feature representations via autoencoders and jointly optimizes a clustering objective; ClusterGAN [Mukherjee *et al.*, 2019] implements clustering through adversarial learning. Subsequently, self-supervised approaches were developed: DeepCluster [Caron *et al.*, 2018] iteratively uses clustering outcomes as pseudo-labels for training, and CC [Li *et al.*, 2021b] further integrates contrastive learning with clustering objectives to enhance performance.

Vision-language pre-training models such as CLIP [Radford *et al.*, 2021] have shifted image clustering from relying on internal priors to exploiting external semantic information. Based on CLIP, SIC [Cai *et al.*, 2023] and TAC [Li *et al.*, 2023b] enhance clustering by modeling image-text geometric consistency and incorporating WordNet semantics with cross-modal distillation, respectively. GradNorm [Peng *et al.*, 2025] further improves text-assisted image clustering by selecting semantically relevant nouns from unlabeled data. [Wang *et al.*, 2026] extend text-guided learning to visual adaptation through deep manifold constraints and neighborhood propagation.

Despite these advancements, existing multimodal clustering methods remain fundamentally limited by their reliance on matching mechanisms within a constrained vocabulary space. Due to the restricted expressiveness of predefined vocabularies, forced alignment between images and textual labels inevitably leads to semantic inconsistency when complex visual semantics exceed the coverage of the fixed lexical set.

Moreover, existing strategies that incorporate semantic priors primarily target bias reduction to improve clustering. It is important to note that estimation error stems from the combined effects of bias, variance, and noise. Therefore, to achieve more robust and stable clustering, addressing variance reduction is also essential. Ensemble learning serves as a key technique for variance reduction in machine learning.

2.2 Ensemble Learning

Ensemble learning addresses complex tasks by constructing and combining multiple learners, and has been widely studied as an effective strategy for improving robustness and stability through the integration of complementary results.

In clustering, different initializations, data views, or similarity measures may produce inconsistent partitions, making the final results sensitive to randomness and data perturbations. To alleviate this issue, representative studies have explored multigranulation information fusion [Li *et al.*, 2017], sample-stability-based clustering ensemble [Li *et al.*, 2019], cluster quality evaluation and selective clustering ensemble [Li *et al.*, 2018], and space-structure-based clustering for mixed-type data [Li *et al.*, 2022]. Reliable evaluation has also been investigated by mitigating random consistency in accuracy measures [Wang *et al.*, 2020] and analyzing the generalization performance of pure accuracy for selective ensemble learning [Wang *et al.*, 2022]. Recent methods further enhance clustering ensemble by designing structured fusion mechanisms, including the growing tree model GOT [Li *et al.*, 2021a], fuzzy ensemble clustering with self-coassociation and prototype propagation [Li *et al.*, 2023a], and k-HyperEdge medoids [Li *et al.*, 2025a].

These studies show that ensemble learning can exploit complementary information from multiple learners, partitions, samples, clusters, or semantic structures to produce more stable and reliable results. By aggregating diverse yet correlated outputs, ensemble methods can effectively suppress the fluctuation of individual predictions and thereby reduce the variance of the final decision.

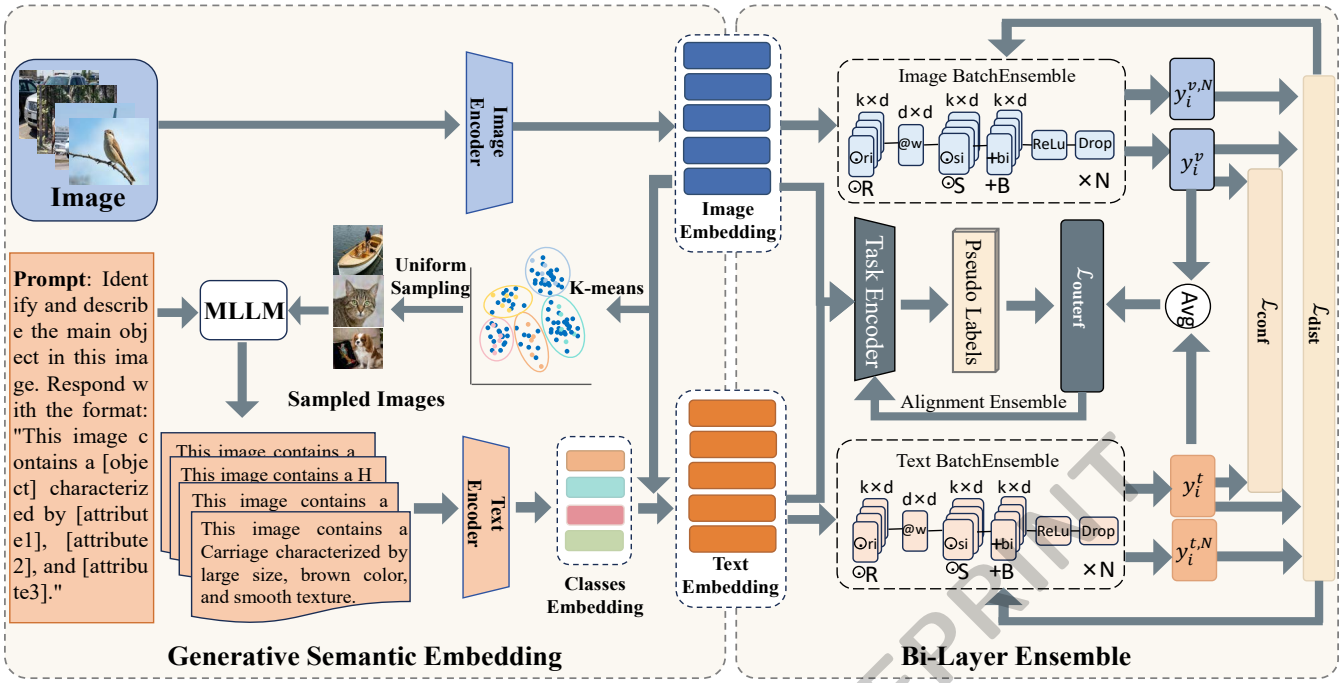

Figure 1: Overall Framework of GSEC. The framework integrates generative semantic embedding with a bi-layer ensemble strategy.

3 Method

In this section, we propose an image clustering method based on generative semantic guidance and bi-Layer ensemble (GSEC), with its overall framework illustrated in Figure 1. Firstly, we introduce the generation of semantic embeddings using Multimodal Large Language Models. Secondly, we elaborate on the operational mechanism of the designed bi-layer ensemble framework.

3.1 Generative Semantic Embedding

To address the limitation of constrained semantic matching space, we leverage Multimodal Large Language Models (MLLMs) to generate semantically enriched descriptions through reasoning. The procedure is detailed as follows.

Let $\mathcal{V} = \{I_i\}_{i=1}^n$ denote the image set, where n is the number of samples. Using an image encoder E_v, we first extract image embeddings $v_i = E_v(I_i) \in \mathbb{R}^d$, with d representing the embedding dimensionality.

To uncover latent semantic structures, we perform K-means clustering on all image embeddings $\{v_i\}_{i=1}^n$. Following the practice in [Peng *et al.*, 2025; Li *et al.*, 2023b], the number of clusters is set as $C = \max\left\{\frac{n}{300}, 3K\right\}$, where K denotes the expected number of clusters. Images within each cluster are then sorted by the distance between their embeddings and the corresponding cluster center. From each cluster, we uniformly select up to five representative images according to this sorted order. If a cluster contains fewer than five images, all of them are selected.

These selected images are fed into the Multimodal Large Language Model using a natural-language prompt designed to capture core visual semantics. The prompt is:" Identify and describe the main object in this image. Respond with the format:' This image contains a [object] characterized by [attribute1], [attribute2], and [attribute3]'".

The generated class-level descriptions T_k are encoded by the CLIP text encoder E_t into class embeddings $\bar{t}_j = E_t(T_k)$. The similarity between each image embedding v_i and every class text embedding $\{\bar{t}_j\}_{j=1}^M$ is then computed to derive an image-specific text embedding t_i, where M is the number of selected samples.

Specifically, we first compute the conditional probability of the j-th class text embedding given image v_i:

$$p(\bar{t}_j|v_i) = \frac{\exp(\text{sim}(v_i, \bar{t}_j)/\tilde{\tau})}{\sum_{k=1}^M \exp(\text{sim}(v_i, \bar{t}_k)/\tilde{\tau})}, \tag{1}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, and $\tilde{\tau}$ is a temperature parameter that controls the smoothness of the distribution. Subsequently, the final text counterpart t_i is obtained as a weighted average of all classes embedding:

$$t_i = \sum_{j=1}^M p(\bar{t}_j|v_i)\bar{t}_j. \tag{2}$$

Here, t_i represents the text embedding of the i-th sample.

3.2 Bi-Layer Ensemble

To mitigate variance, we design a bi-layer ensemble framework: the inner layer employs a BatchEnsemble integrator for cross-modal clustering, and the outer layer further aligns and integrates the inner outputs with external predictions through an alignment-based ensemble mechanism.

Inner Ensemble

At the inner layer, we construct an integrator based on the BatchEnsemble method to efficiently integrate cross-modal data. This integrator consists of two independent branches for image and text modalities, each comprising a set of shared weights and member-specific low-rank modulation parameters.

The output of the k-th base model in the BatchEnsemble integrator is given by:

$$l_k(x) = s_k \odot (W(r_k \odot x)) + b_k, \tag{3}$$

where W denotes the shared weight matrix across all members, r_k and s_k are learnable rank-1 vectors specific to the k-th member, $\odot$ represents the element-wise multiplication, and b_k is the bias term for the k-th member.

We denote the BatchEnsemble integrators for the image and text modalities as $l_k^v(x)$ and $l_k^t(x)$, respectively. Let v_i^N be a random neighbor of v_i, and t_i^N be a random neighbor of t_i. We process v_i and v_i^N through the image BatchEnsemble integrator, and t_i and t_i^N through the text BatchEnsemble integrator. As for y_i^v and y_i^t:

$$y_i^v = \frac{1}{m}\sum_{k=1}^{m} l_k^v(v_i), \quad y_i^t = \frac{1}{m}\sum_{k=1}^{m} l_k^t(t_i), \tag{4}$$

where m is the ensemble size, y_i^v and y_i^t represent the soft assignments for the i-th image and text, respectively. As for $y_i^{v,N}$ and $y_i^{t,N}$:

$$y_i^{v,N} = \frac{1}{m}\sum_{k=1}^{m} l_k^v(v_i^N), \quad y_i^{t,N} = \frac{1}{m}\sum_{k=1}^{m} l_k^t(t_i^N), \tag{5}$$

where $y_i^{v,N}$ and $y_i^{t,N}$ denote the neighbor soft assignments for the i-th image and text, respectively.

To achieve mutual guidance between the text and image modalities during clustering, we introduce the Cross-Modal Mutual Distillation loss:

$$\mathcal{L}_{\text{dist}} = \sum_{i=1}^{n}\left(D_{\text{KL}}(y_i^t \| y_i^{v,N}) + D_{\text{KL}}(y_i^v \| y_i^{t,N})\right), \tag{6}$$

where D_{KL} denotes the KL divergence. This loss serves a dual purpose: it enhances the discriminability of cluster boundaries by reducing inter-cluster similarity, while also enabling bidirectional alignment between each image and the neighbors of its corresponding text, and between each text and the neighbors of its corresponding image.

To stabilize training and encourage the model to produce more confident cluster assignments, we introduce the confidence loss:

$$\mathcal{L}_{\text{conf}} = -\log \sum_{i=1}^{n} (y_i^v)^\top y_i^t. \tag{7}$$

To prevent the collapse of all samples into only a few clusters, we adopt the balance loss:

$$\mathcal{L}_{\text{bal}} = -\sum_{i=1}^{K}\left(\bar{y^v}\log \bar{y^v} + \bar{y^t}\log \bar{y^t}\right), \tag{8}$$

where K is the number of clusters. $\bar{y}^v$ and $\bar{y}^t$ denote the cluster assignment distributions in the image and text modalities, respectively, which are computed as:

$$\bar{y^t} = \frac{1}{n}\sum_{i=1}^{n} y_i^t, \quad \bar{y^v} = \frac{1}{n}\sum_{i=1}^{n} y_i^v. \tag{9}$$

Then, the inner ensemble aims to optimize the model parameters by minimizing the following comprehensive loss function:

$$\mathcal{L}_{\text{inner}} = \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{conf}} - \mathcal{L}_{\text{bal}}. \tag{10}$$

Outer Ensemble

After the inner-layer ensemble training converges, we compute the average of the outputs from the image and text modalities:

$$\hat{y}_i = \frac{1}{2}(y_i^v + y_i^t), \tag{11}$$

where $\hat{y}_i$ denotes the output of the inner ensemble.

In the outer ensemble stage, we train a task encoder ϕ that takes the concatenated image and text embeddings $[v_i; t_i]$ as input, with its output defined as:

$$y_i = \phi\left([v_i; t_i]\right). \tag{12}$$

To align the predictions of the inner ensemble with the outputs of the outer task encoder, we adopt the cross-entropy loss as the alignment objective:

$$\mathcal{L}_{\text{align}} = \sum_{i=1}^{n} \mathcal{L}_{\text{CE}}(y_i, \hat{y}_i). \tag{13}$$

The outer ensemble optimizes the parameters ϕ by minimizing the following comprehensive loss:

$$\mathcal{L}_{\text{outer}} = \mathcal{L}_{\text{align}} - H(\bar{p}), \tag{14}$$

where $\bar{p} = \frac{1}{n}\sum_{i=1}^{n}\phi([\mathbf{v}_i; \mathbf{t}_i])$ is the average prediction distribution, and $H(\bar{p})$ denotes its entropy.

After the outer ensemble training converges, the final cluster assignment for each sample is obtained from the output of the task encoder. This decision mechanism integrates the multimodal alignment information from the inner ensemble outputs with the comprehensive inference of the task encoder, which is expected to generate a stable and consistent clustering result.

4 Experiments

In this section, we validate GSEC's performance through benchmark comparisons. Additionally, we conduct bias-variance decomposition analysis to verify the effectiveness of our method, along with ablation studies, hyperparameter analysis, convergence analysis, practicality analysis, and visualization analysis (In Appendix [Li *et al.*, 2026]).

4.1 Experiment Setup

Datasets To evaluate the effectiveness of our proposed method, we conduct experiments on 11 widely used benchmark datasets. The detailed information of these datasets is summarized in Table 1.

No	Dataset	Cls	Train size	Test size
1	CIFAR10 [Krizhevsky *et al.*, 2009]	10	50,000	10,000
2	CIFAR100 [Krizhevsky *et al.*, 2009]	100	50,000	10,000
3	STL10 [Coates *et al.*, 2011]	10	5,000	8,000
4	ImageNet-10 [Chang *et al.*, 2017]	10	13,000	500
5	ImageNet-Dogs [Chang *et al.*, 2017]	15	19,500	750
6	Food101 [Bossard *et al.*, 2014]	101	75,750	25,250
7	StanfordCars [Krause *et al.*, 2013]	196	8,144	8,041
8	OxfordPets [Parkhi *et al.*, 2012]	37	3,680	3,669
9	FGVC Aircraft [Maji *et al.*, 2013]	100	6,667	3,333
10	Country211 [Radford *et al.*, 2021]	211	42,200	21,100
11	ImageNet-1K [Deng *et al.*, 2009]	1000	1,281,167	50,000

Table 1: Statistics of the benchmark datasets. This table summarizes the details of the eleven datasets used in the experiments.

Evaluation To evaluate the clustering performance, we employ three widely adopted metrics: Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI). For all three metrics, higher values indicate better performance.

Implementation Details We utilize the pre-trained CLIP-ViT-B/32 model to extract image and text embeddings. For semantic descriptions, we employ Llama-3.2-Vision-11B to generate descriptions from representative images, which are then standardized into a unified format. The learning rates for both the internal integrator and the task encoder are set to 0.001, and the number of ensemble is set to 24. The batch size is fixed at 1024, with the exception of ImageNet-1K, which uses a batch size of 8192. All experiments were conducted on a single NVIDIA A6000 GPU.

4.2 Comparison with Baseline Methods

To comprehensively assess the effectiveness of GSEC, we benchmark it against a wide range of state-of-the-art algorithms, categorized into two groups: (1) Unimodal Clustering Methods, including DEC [Xie *et al.*, 2016], DAC [Chang *et al.*, 2017], JULE [Yang *et al.*, 2016], DCCM [Wu *et al.*, 2019], IIC [Ji *et al.*, 2019], SCAN [Van Gansbeke *et al.*, 2020], TCL [Tao *et al.*, 2021], TCC [Shen *et al.*, 2021], GCC [Zhong *et al.*, 2021], CRLC [Do *et al.*, 2021], RPSC [Liu *et al.*, 2024b], PICA [Huang *et al.*, 2020], SPICE [Niu *et al.*, 2022], DivClust [Metaxas *et al.*, 2023], and LFSS [Li *et al.*, 2025b]; (2) Multimodal Clustering Methods, such as TAC [Li *et al.*, 2023b], CLIP (K-means), SIC [Cai *et al.*, 2023], GradNorm [Peng *et al.*, 2025], and CAE [Zhu *et al.*, 2026].

We evaluate our proposed GSEC on the first five datasets listed in Table 1 and compare it with the above mentioned 18 image clustering baselines. As shown in Table 2, our method consistently outperforms these competitors. Furthermore, we test GSEC on the challenging large-scale ImageNet-1K dataset and compare it with four leading multimodal methods. As presented in Table 5, our approach retains its superiority and delivers the best performance.

4.3 Reduction of Bias-Variance Analysis

To evaluate the effectiveness of the proposed method in reducing bias and variance, we first generate 10 base datasets via resampling with the same sample size for each dataset. Each of these datasets is used to train a model. Bias is defined as the average deviation between the model's predictions and the ground truth, while variance quantifies the dispersion across predictions from individual models.

We test the bias-variance performance of the following five configurations. **Image**: using only image features with the bi-layer linear architecture; **Image+Ensemble**: using only image features with the bi-layer ensemble framework; **Image+M-Text**: feeding image features and matching-based textual features into a bi-layer linear architecture; **Image+G-Text**: feeding image features and generative textual features into the bi-layer linear architecture; **GSEC**: feeding image features and generative textual features into the bi-layer ensemble framework.

The experimental results are illustrated in Figure 2 and Figure 6 (in Appendix [Li *et al.*, 2026]). The following observations can be observed: (1) Comparing **Image** and **Image+Ensemble**, the ensemble strategy effectively reduces variance. (2) Comparing **Image+M-Text** and **Image+G-Text**, generative textual guidance helps reduce bias but introduces additional variance. (3) Compared with other methods, **GSEC** simultaneously reduces bias and variance through generative textual guidance and ensemble strategy, yielding superior and more stable clustering performance.

4.4 Ablation Analysis

To verify the effectiveness of each component in the proposed method, we conduct ablation studies to examine the impact of generative semantic embeddings and the bi-layer ensemble design on the performance of GSEC. Experiments are carried out across multiple pre-trained backbones, including CLIP-ViT-B/32, CLIP-RN50x4, CLIP-RN101, and CLIP-RN50. Performance is evaluated using Accuracy.

Effectiveness of the Generative Semantic We derive semantic prior information from two distinct sources: the Llama-3.2-Vision-11B MLLM for generative semantics (denoted as G-Text) and the WordNet for matching semantics (denoted as M-Text). These textual descriptions are encoded by separate feature extractors and subsequently processed through the same bi-layer ensemble framework for clustering. As shown in Table 3, G-text consistently outperforms M-text across all tested backbones.

Effectiveness of the Bi-layer Ensemble We compare the proposed bi-layer ensemble framework with a bi-layer linear architecture under identical image feature inputs. As reported in Table 4, our ensemble-based approach yields consistent performance gains on ten benchmark datasets across the four different backbones.

4.5 Hyperparameter Analysis

Analysis of the number of ensembles. We analyze the impact of the ensemble size on clustering performance and computational efficiency. As illustrated in Figure 3, the runtime

Dataset	CIFAR10			CIFAR100			ImageNet-10			ImageNet-Dogs			STL10		
Metrics	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
JULE (CVPR16)	19.2	27.2	13.8	10.3	13.7	3.3	17.5	30.0	13.8	5.4	13.8	2.8	18.2	27.7	16.4
DEC (ICML16)	25.7	30.1	16.1	13.6	18.5	5.0	28.2	38.1	20.3	12.2	19.5	7.9	27.6	35.9	18.6
DAC (ICCV17)	39.6	52.2	30.6	18.5	23.8	8.8	39.4	52.7	30.2	21.9	27.5	11.1	36.6	47.0	25.7
DCCM (ICCV19)	49.6	62.3	40.8	28.5	32.7	17.3	60.8	71.0	55.5	32.1	38.3	18.2	37.6	48.2	26.2
IIC (ICCV19)	51.3	61.7	41.1	22.5	25.7	11.7	–	–	–	–	–	–	49.6	59.6	39.7
PICA (CVPR20)	59.1	69.6	51.2	31.0	33.7	17.1	80.2	87.0	76.1	35.2	35.3	20.1	61.1	71.3	53.1
SCAN (ECCV20)	79.7	88.3	77.2	48.6	50.7	33.3	–	–	–	61.2	59.3	45.7	69.8	80.9	64.6
GCC (ICCV21)	76.4	85.6	72.8	47.2	47.2	30.5	84.2	90.1	82.2	49.0	52.6	36.2	68.4	78.8	63.1
CRLC (ICCV21)	67.9	79.9	63.4	41.6	42.5	26.3	83.1	85.4	75.9	48.4	46.1	59.7	72.9	81.8	62.8
TCC (NeurIPS21)	79.0	90.6	73.3	47.9	49.1	31.2	84.8	89.7	82.5	55.4	59.5	41.7	72.2	81.4	68.9
TCL (IJCV22)	81.9	88.7	78.0	52.9	53.1	35.7	87.5	89.5	83.7	62.3	64.4	51.6	79.9	86.8	75.7
SPICE (TIP22)	73.4	83.8	70.5	44.8	46.8	29.4	82.8	92.1	83.6	57.2	64.6	68.7	81.7	90.8	81.2
DivClust (CVPR23)	71.0	81.5	67.5	44.0	43.7	28.3	85.0	90.0	81.9	51.6	52.9	37.6	–	–	–
RPSC (AAAI24)	75.4	85.7	73.1	47.6	51.8	34.1	83.0	92.7	85.8	55.2	64.0	46.5	83.8	92.0	83.4
LFSS (ICML25)	87.2	93.4	86.8	–	–	–	85.6	93.2	85.7	61.6	69.1	53.3	77.1	86.1	74.0
CLIP (k-means)	70.3	74.2	61.6	59.7	43.4	29.1	96.9	98.2	96.1	39.8	38.1	20.1	91.7	94.3	89.1
SIC (AAAI23)	84.7	92.6	84.4	63.2	47.6	34.8	97.0	98.2	96.1	69.0	69.7	55.8	95.3	98.1	95.9
TAC (ICML24)	83.3	91.9	83.1	67.5	56.1	40.8	98.5	99.2	98.3	80.6	83.0	72.2	95.5	98.2	96.1
GradNorm (IEEE25)	82.6	91.1	81.5	67.2	54.8	37.0	98.7	99.4	98.7	81.0	81.2	70.9	95.7	98.3	96.2
CAE (AAAI26)	81.7	90.9	80.5	–	–	–	98.1	98.8	97.4	77.9	77.5	66.3	95.5	98.2	96.7
GSEC	**91.2**	**96.3**	**92.1**	**71.8**	**61.8**	**48.0**	**99.6**	**99.8**	**99.6**	**86.3**	**87.1**	**79.2**	**97.6**	**99.4**	**97.7**

Table 2: Comparison with reference baseline methods. The best results are highlighted in bold.

Backbone	CIFAR10			CIFAR100			ImageNet-Dogs			STL10			ImageNet-10		
	M-Text	G-Text	Gain	M-Text	G-Text	Gain	M-Text	G-Text	Gain	M-Text	G-Text	Gain	M-Text	G-Text	Gain
CLIP-RN50	71.56	72.10	**0.54**	31.59	34.87	**3.28**	77.82	78.80	**0.98**	61.46	96.70	**35.24**	98.40	98.52	**0.12**
CLIP-RN101	81.74	81.83	**0.09**	40.47	44.50	**4.03**	76.93	82.93	**6.00**	76.71	98.11	**21.40**	98.60	98.63	**0.03**
CLIP-RN50x4	76.39	78.24	**1.85**	38.41	38.65	**0.24**	82.40	83.33	**0.93**	70.35	97.92	**27.57**	98.20	98.80	**0.60**
CLIP-ViT-B/32	89.58	90.16	**0.58**	46.41	49.34	**2.93**	77.20	78.67	**1.47**	73.90	98.32	**24.42**	98.20	98.72	**0.52**

Backbone	StanfordCars			OxfordPets			Aircraft			Country211			Food101		
	M-Text	G-Text	Gain	M-Text	G-Text	Gain	M-Text	G-Text	Gain	M-Text	G-Text	Gain	M-Text	G-Text	Gain
CLIP-RN50	30.52	50.22	**19.70**	59.42	79.50	**20.08**	19.20	24.42	**5.22**	9.12	12.37	**3.25**	54.54	74.96	**20.42**
CLIP-RN101	28.04	56.22	**28.18**	62.20	82.42	**20.22**	22.26	27.62	**5.36**	9.65	13.67	**4.02**	63.18	78.38	**15.20**
CLIP-RN50x4	34.57	61.45	**26.88**	70.37	87.49	**17.12**	25.92	30.40	**4.48**	10.41	15.36	**4.95**	67.30	78.77	**11.47**
CLIP-ViT-B/32	37.33	50.69	**13.36**	61.52	73.83	**12.31**	15.99	26.53	**10.54**	9.22	13.06	**3.84**	66.90	78.22	**11.32**

Table 3: Effectiveness Analysis of Semantic Prior Information. The table reports the clustering accuracy achieved using semantic priors obtained from WordNet matching (M-Text) and generative descriptions (G-Text).

Backbone	CIFAR10			CIFAR100			ImageNet-Dogs			STL10			ImageNet-10		
	Image	Ensemble	Gain	Image	Ensemble	Gain	Image	Ensemble	Gain	Image	Ensemble	Gain	Image	Ensemble	Gain
CLIP-RN50	57.20	62.44	**5.24**	30.80	31.23	**0.43**	45.20	47.07	**1.87**	96.60	96.69	**0.09**	98.20	98.40	**0.20**
CLIP-RN101	71.60	74.68	**3.08**	40.20	41.42	**1.22**	51.13	66.27	**15.14**	98.00	98.02	**0.02**	98.20	98.60	**0.40**
CLIP-RN50x4	67.70	69.90	**2.20**	37.70	37.83	**0.13**	59.47	74.13	**14.66**	97.80	97.89	**0.09**	98.60	98.80	**0.20**
CLIP-ViT-B/32	85.90	86.19	**0.29**	45.80	46.61	**0.81**	48.40	63.60	**15.20**	98.30	98.31	**0.01**	98.40	98.60	**0.20**

Backbone	StanfordCars			OxfordPets			Aircraft			Country211			Food101		
	Image	Ensemble	Gain	Image	Ensemble	Gain	Image	Ensemble	Gain	Image	Ensemble	Gain	Image	Ensemble	Gain
CLIP-RN50	37.00	38.85	**1.85**	51.20	54.10	**2.90**	23.00	24.06	**1.06**	8.90	9.39	**0.49**	65.00	67.73	**2.73**
CLIP-RN101	46.30	49.67	**3.37**	66.70	71.49	**4.79**	24.20	27.36	**3.16**	9.50	9.65	**0.15**	74.90	75.21	**0.31**
CLIP-RN50x4	52.90	53.23	**0.33**	74.30	74.65	**0.35**	26.70	30.27	**3.57**	10.10	10.41	**0.31**	80.70	81.24	**0.54**
CLIP-ViT-B/32	42.30	44.09	**1.79**	62.10	63.91	**1.81**	24.00	26.01	**2.01**	9.80	9.83	**0.03**	72.40	73.78	**1.38**

Table 4: Effectiveness of the bi-layer ensemble strategy across different backbones. The table compares the clustering accuracy of the single-image baseline (Image) and the ensemble-based method (Ensemble) on 10 downstream tasks.

increases with the number of ensemble members, and an ensemble size of 24 achieves a favorable balance between performance gains and computational cost.

Analysis of Internal and External Learning Rates. We investigate the effect of different combinations of inner and outer learning rates on clustering performance, aiming to identify an optimal configuration that ensures stable clustering accuracy. As shown in Figure 4, the model attains the best overall performance across all ten benchmark datasets when both the inner and outer learning rates are set to 0.001.

4.6 Convergence Analysis

To examine the training convergence of GSEC, we visualize the evolution of both inner and outer losses over training

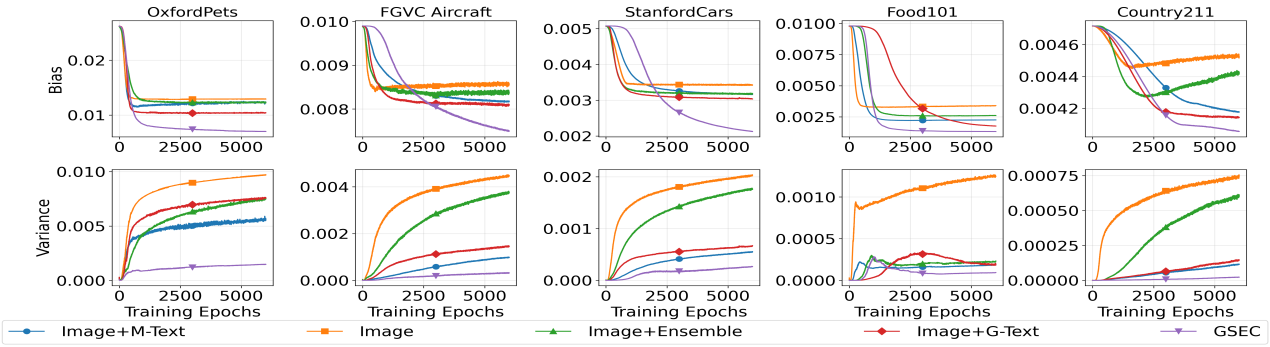

Figure 2: Bias–Variance Analysis. The figure visualizes the evolution of bias (top row) and variance (bottom row) across five datasets. The curves illustrate that GSEC (purple) consistently achieves the lowest bias and variance compared to other variants.

Metric	CLIP (k-means)	SIC	TAC	GradNorm	**GSEC**
NMI	72.3	77.2	77.8	79.2	**81.3**
ACC	38.9	47.0	48.9	52.6	**62.5**
ARI	27.1	34.3	36.4	39.1	**52.8**

Table 5: Comparison on the large-scale ImageNet-1K dataset.

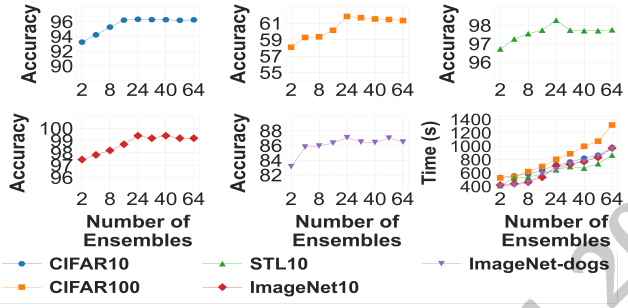

Figure 3: Sensitivity analysis of the ensemble size. The figure evaluates the influence of the number of ensemble members on clustering accuracy (first five subplots) and computational time (last subplot) across five benchmark datasets.

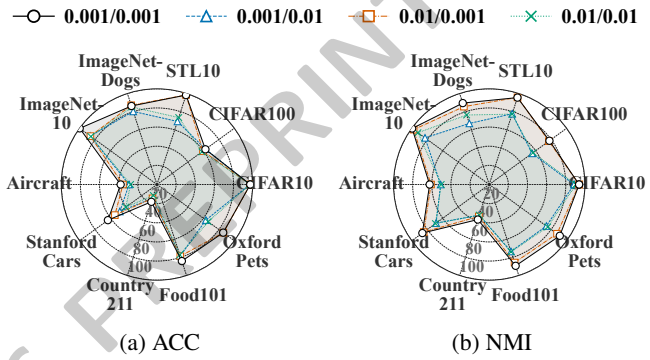

Figure 4: Sensitivity analysis of learning rates. The radar chart illustrates the clustering performance across ten benchmark datasets under four different combinations of inner and outer learning rates.

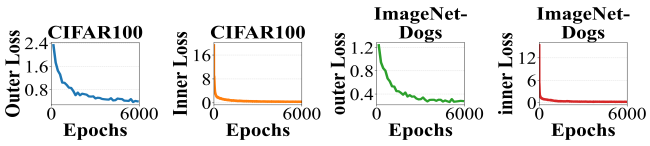

Figure 5: Convergence analysis of GSEC. The figure shows the training curves of the outer and inner losses on CIFAR-100 and ImageNet-Dogs datasets, illustrating stable optimization behavior.

epochs. As shown in Figure 5, both losses exhibit a clear downward trend and eventually converge, demonstrating stable optimization.

4.7 Practicality Analysis

To assess the efficiency of semantic querying, we report the total query tokens and average latency on CIFAR10 and STL10. As shown in Table 6, our method incurs limited query overhead and achieves practical response efficiency, with latency of only a few seconds.

CIFAR10	Query (Tokens)	Latency (S)	STL10	Query (Tokens)	Latency (S)
	428986	4.07		48036	4.80

Table 6: Evaluation of Query Cost and Average Latency.

5 Conclusion

This paper proposes GSEC (Image Clustering based on Generative Semantic Guidance and Bi-Layer Ensemble), a novel framework that jointly reduces bias and variance in image clustering. To overcome the limitations of vocabulary-matching-based semantics, GSEC leverages Multimodal Large Language Models to generate adaptive semantic descriptions, which mitigate semantic inconsistency and reduce bias. Simultaneously, a bi-layer ensemble mechanism is introduced to lower variance and improve robustness. This ensemble consists of an inner layer that integrates cross-modal information via BatchEnsemble and an outer layer that aligns inner and outer predictions through an alignment-based ensemble. Experiments on 6 benchmark datasets demonstrate the effectiveness of GSEC in image clustering. Further bias-variance decomposition confirms that GSEC reduces both bias and variance concurrently, highlighting the importance of jointly optimizing these two factors for improved clustering performance.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62441239, 62476160, T2495251, 62441239, 62306170, U24A20253, 62136005), the Special Fund for Science and Technology Innovation Teams of Shanxi Province (No. 202304051001001).

References

[Bossard *et al.*, 2014] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014.

[Cai *et al.*, 2023] Shaotian Cai, Liping Qiu, Xiaojun Chen, Qin Zhang, and Longteng Chen. Semantic-enhanced image clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 6869–6878, 2023.

[Caron *et al.*, 2018] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018.

[Chang *et al.*, 2017] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *Proceedings of the IEEE international conference on computer vision*, pages 5879–5887, 2017.

[Coates *et al.*, 2011] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[Dasgupta, 2002] Sanjoy Dasgupta. Performance guarantees for hierarchical clustering. In *International Conference on Computational Learning Theory*, pages 351–363. Springer, 2002.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[Do *et al.*, 2021] Kien Do, Truyen Tran, and Svetha Venkatesh. Clustering by maximizing mutual information across views. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9928–9938, 2021.

[Ester *et al.*, 1996] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[Huang *et al.*, 2020] Jiabo Huang, Shaogang Gong, and Xiatian Zhu. Deep semantic clustering by partition confidence maximisation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8849–8858, 2020.

[Ji *et al.*, 2019] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9865–9874, 2019.

[Krause *et al.*, 2013] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

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

[Li *et al.*, 2017] Feijiang Li, Yuhua Qian, Jieting Wang, and Jiye Liang. Multigranulation information fusion: A dempster-shafer evidence theory-based clustering ensemble method. *Information Sciences*, 378:389–409, 2017.

[Li *et al.*, 2018] Feijiang Li, Yuhua Qian, Jieting Wang, Chuangyin Dang, and Bing Liu. Cluster's quality evaluation and selective clustering ensemble. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 12(5):1–27, 2018.

[Li *et al.*, 2019] Feijiang Li, Yuhua Qian, Jieting Wang, Chuangyin Dang, and Liping Jing. Clustering ensemble based on sample's stability. *Artificial Intelligence*, 273:37–55, 2019.

[Li *et al.*, 2021a] Feijiang Li, Yuhua Qian, and Jieting Wang. Got: a growing tree model for clustering ensemble. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 8349–8356, 2021.

[Li *et al.*, 2021b] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 8547–8555, 2021.

[Li *et al.*, 2022] Feijiang Li, Yuhua Qian, Jieting Wang, Furong Peng, and Jiye Liang. Clustering mixed type data: A space structure-based approach. *International Journal of Machine Learning and Cybernetics*, 13(9):2799–2812, 2022.

[Li *et al.*, 2023a] Feijiang Li, Jieting Wang, Yuhua Qian, Guoqing Liu, and Keqi Wang. Fuzzy ensemble clustering based on self-coassociation and prototype propagation. *IEEE Transactions on Fuzzy Systems*, 31(10):3610–3623, 2023.

[Li *et al.*, 2023b] Yunfan Li, Peng Hu, Dezhong Peng, Jiancheng Lv, Jianping Fan, and Xi Peng. Image clustering with external guidance. *arXiv preprint arXiv:2310.11989*, 2023.

[Li *et al.*, 2025a] Feijiang Li, Jieting Wang, Liuya Zhang, Yuhua Qian, Shuai Jin, Tao Yan, and Liang Du. k-hyperedge medoids for clustering ensemble. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18271–18278, 2025.

[Li *et al.*, 2025b] Zhixin Li, Yuheng Jia, Junhui Hou, et al. Learning from sample stability for deep clustering.

In *Forty-second International Conference on Machine Learning*, 2025.

[Li *et al.*, 2026] Feijiang Li, Zhenxiong Li, Jieting Wang, Zizheng Jiu, Saixiong Liu, and Liang Du. Reducing bias and variance: Generative semantic guidance and bi-layer ensemble for image clustering, 2026.

[Liu *et al.*, 2024a] Honglin Liu, Peng Hu, Changqing Zhang, Yunfan Li, and Xi Peng. Interactive deep clustering via value mining. *Advances in Neural Information Processing Systems*, 37:42369–42387, 2024.

[Liu *et al.*, 2024b] Sihang Liu, Wenming Cao, Ruigang Fu, Kaixiang Yang, and Zhiwen Yu. Rpsc: Robust pseudo-labeling for semantic clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14008–14016, 2024.

[Maji *et al.*, 2013] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[McQueen, 1967] James B McQueen. Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, pages 281–297, 1967.

[Metaxas *et al.*, 2023] Ioannis Maniadis Metaxas, Georgios Tzimiropoulos, and Ioannis Patras. Divclust: Controlling diversity in deep clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3418–3428, 2023.

[Mukherjee *et al.*, 2019] Sudipto Mukherjee, Himanshu Asnani, Eugene Lin, and Sreeram Kannan. Clustergan: Latent space clustering in generative adversarial networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4610–4617, 2019.

[Niu *et al.*, 2022] Chuang Niu, Hongming Shan, and Ge Wang. Spice: Semantic pseudo-labeling for image clustering. *IEEE Transactions on Image Processing*, 31:7264–7278, 2022.

[Parkhi *et al.*, 2012] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

[Peng *et al.*, 2025] Bo Peng, Jie Lu, Guangquan Zhang, and Zhen Fang. On the provable importance of gradients for autonomous language-assisted image clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19805–19815, 2025.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[Ren *et al.*, 2024] Yazhou Ren, Jingyu Pu, Zhimeng Yang, Jie Xu, Guofeng Li, Xiaorong Pu, Philip S Yu, and Lifang He. Deep clustering: A comprehensive survey. *IEEE transactions on neural networks and learning systems*, 36(4):5858–5878, 2024.

[Shen *et al.*, 2021] Yuming Shen, Ziyi Shen, Menghan Wang, Jie Qin, Philip Torr, and Ling Shao. You never cluster alone. *Advances in Neural Information Processing Systems*, 34:27734–27746, 2021.

[Tao *et al.*, 2021] Yaling Tao, Kentaro Takagi, and Kouta Nakata. Clustering-friendly representation learning via instance discrimination and feature decorrelation. *arXiv preprint arXiv:2106.00131*, 2021.

[Van Gansbeke *et al.*, 2020] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European conference on computer vision*, pages 268–285. Springer, 2020.

[Wang *et al.*, 2020] Jieting Wang, Yuhua Qian, and Feijiang Li. Learning with mitigating random consistency from the accuracy measure. *Machine Learning*, 109(12):2247–2281, 2020.

[Wang *et al.*, 2022] Jieting Wang, Yuhua Qian, Feijiang Li, Jiye Liang, and Qingfu Zhang. Generalization performance of pure accuracy and its application in selective ensemble learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1798–1816, 2022.

[Wang *et al.*, 2026] Jieting Wang, Yue Sun, Feijiang Li, and Huimei Shi. Text-guided domain adaptation via deep manifold constraints and neighborhood propagation. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2586–2590. IEEE, 2026.

[Wu *et al.*, 2019] Jianlong Wu, Keyu Long, Fei Wang, Chen Qian, Cheng Li, Zhouchen Lin, and Hongbin Zha. Deep comprehensive correlation mining for image clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8150–8159, 2019.

[Xie *et al.*, 2016] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016.

[Yang *et al.*, 2016] Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5147–5156, 2016.

[Zhong *et al.*, 2021] Huasong Zhong, Jianlong Wu, Chong Chen, Jianqiang Huang, Minghua Deng, Liqiang Nie, Zhouchen Lin, and Xian-Sheng Hua. Graph contrastive clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9224–9233, 2021.

[Zhu *et al.*, 2026] Xingyu Zhu, Beier Zhu, Yunfan Li, Junfeng Fang, Shuo Wang, Kesen Zhao, and Hanwang Zhang. Hierarchical semantic alignment for image clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 29177–29185, 2026.