Image Clustering with External Guidance

Yunfan Li [1] **Peng Hu** [1] **Dezhong Peng** [1] **Jiancheng Lv** [1] **Jianping Fan** [2] **Xi Peng** [1]

Abstract

The core of clustering lies in incorporating prior knowledge to construct supervision signals. From classic k-means based on data compactness to recent contrastive clustering guided by self-supervision, the evolution of clustering methods intrinsically corresponds to the progression of supervision signals. At present, substantial efforts have been devoted to mining internal supervision signals from data. Nevertheless, the abundant external knowledge such as semantic descriptions, which naturally conduces to clustering, is regrettably overlooked. In this work, we propose leveraging external knowledge as a new supervision signal to guide clustering. To implement and validate our idea, we design an externally guided clustering method (Text-Aided Clustering, TAC), which leverages the textual semantics of WordNet to facilitate image clustering. Specifically, TAC first selects and retrieves WordNet nouns that best distinguish images to enhance the feature discriminability. Then, TAC collaborates text and image modalities by mutually distilling cross-modal neighborhood information. Experiments demonstrate that TAC achieves state-of-the-art performance on five widely used and three more challenging image clustering benchmarks, including the full ImageNet-1K dataset. The code can be accessed at https://github.com/XLearning-SCU/2024-ICML-TAC.
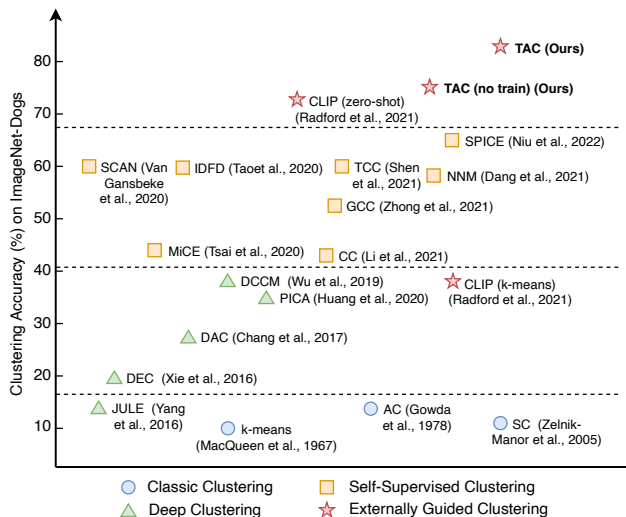
Figure 1. The evolution of clustering methods could be roughly divided into three eras, including **i) classic clustering**, which designs clustering strategies based on data distribution assumptions; **ii) deep clustering**, which extracts clustering-favorable features with deep neural networks, and **iii) self-supervised clustering**, which constructs supervision signals through data augmentations or momentum strategies. In this work, instead of mining the internal supervision, we propose exploring external knowledge to facilitate image clustering. We categorize such a novel paradigm as **iv) externally guided clustering**. By leveraging the semantics in the text modality, our TAC pushes the clustering accuracy to a new state-of-the-art.

1. Introduction

Image clustering aims at partitioning images into different groups in an unsupervised fashion, which is a long-standing task in machine learning. The core of clustering resides in incorporating prior knowledge to construct supervision signals. According to different choices of supervision signals,

[1]School of Computer Science, Sichuan University, Chengdu, China [2]AI Lab at Lenovo Research, Beijing, China. Correspondence to: Xi Peng <pengx.gm@gmail.com>.

one could roughly divide the evolution of clustering methods into three eras, *i.e.*, classic clustering, deep clustering, and self-supervised clustering as depicted in Fig 1. At the early stage, classic clustering methods build upon various assumptions on the data distribution, such as compactness (MacQueen et al., 1967; Ester et al., 1996), hierarchy (Gowda & Krishna, 1978), connectivity (Zelnik-Manor & Perona, 2005; Nie et al., 2011; Wang et al., 2020), sparsity (Elhamifar & Vidal, 2013; Liu et al., 2017), and low rank (Cai et al., 2009; Liu et al., 2012; Nie et al., 2016). Though having achieved promising performance, classic clustering methods would produce suboptimal results confronting complex and high-dimensional data. As an improvement, deep clustering methods equip clustering models with neural networks

to extract discriminative features (Peng et al., 2016; Yang et al., 2016; Xie et al., 2016; Li et al., 2020). In alignment with priors such as cluster discriminability (Ghasedi Dizaji et al., 2017) and balance (Hu et al., 2017), various supervision signals are formulated to optimize the clustering network. In the last few years, motivated by the success of self-supervised learning (He et al., 2020; Chen et al., 2020; Grill et al., 2020), clustering methods turn to creating supervision signals through data augmentation (Li et al., 2021b; Van et al., 2020; Dang et al., 2021b) or momentum strategies (Zhong et al., 2021; Huang et al., 2022).

Though varying in the method design, most existing clustering methods design supervision signals in an internal manner. Despite the remarkable success achieved, the internally guided clustering paradigm faces an inherent limitation. Specifically, the hand-crafted internal supervision signals, even enhanced with data augmentation, are inherently upper-bounded by the limited information in the given data. For example, "Corgi" and "Shiba Inu" dogs are visually similar and are likely to be confused in image clustering. Luckily, beyond the internal signals, we notice there also exists well-established external knowledge that potentially conduces to clustering, while having been regrettably and largely ignored. In the above example, we could better distinguish the images given the external knowledge that "Corgi" have shorter and thicker legs compared with "Shiba Inu" dogs. In short, from different sources or modalities, the external knowledge could serve as promising supervision signals to guide clustering. Compared with exhaustively mining internal supervision signals from data, it would yield twice the effect with half the effort by incorporating rich and readily available external knowledge to guide clustering.

In this work, we propose a simple yet effective externally guided clustering method TAC (Text-Aided Clustering), which clusters images by incorporating external knowledge from the text modality. In the absence of class name priors, there are two challenges in leveraging the textual semantics for image clustering, namely, *i)* how to construct the text space, and *ii)* how to collaborate images and texts for clustering. For the first challenge, ideally, we expect the text counterparts of between-class images to be highly distinguishable so that clustering can be easily achieved. To this end, inspired by the zero-shot classification paradigm in CLIP (Radford et al., 2021), we reversely classify all nouns from WordNet (Miller, 1995) to image semantic centers. Based on the classification confidence, we select the most discriminative nouns for each image center to form the text space and retrieve a text counterpart for each image. Intriguingly, Fig. 2 demonstrates that in certain cases, the retrieved nouns could describe the image semantics, sometimes even better than the manually annotated class names. For the second challenge, we first establish an extremely simple baseline by concatenating the images and text counterparts,

Figure 2. Our observations with two image examples from the ImageNet-Dogs dataset as a showcase. For each example, we show the manually annotated class names and the nouns obtained by the proposed TAC, as well as the zero-shot classification probabilities. From the example, one could arrive at two observations, namely, i) visually similar samples could be better distinguished in the text modality, and ii) manually annotated class names are not always the best semantic description. As shown, zero-shot CLIP falsely classifies both images to the *Blenheim Spaniel* class (probably due to the word *Spaniel*), whereas the nouns obtained by our TAC successfully separate them. Such observations suggest a great opportunity to leverage the external knowledge (hidden in the text modality in this showcase) to facilitate image clustering.

which already significantly enhances the k-means clustering performance without any additional training. For better collaboration, we propose to mutually distill the neighborhood information between the text and image modalities. By additionally training cluster heads, the proposed TAC achieves state-of-the-art performance on five widely used and three more challenging image clustering datasets. Without loss of generality, we evaluate TAC on the pre-trained CLIP model in our experiments, but TAC could adapt to any vision-language pre-trained (VLP) model by design.

The major contributions of this work could be summarized as follows:

- Unlike previous clustering works that exhaustively explore and exploit supervision signals internally, we propose leveraging external knowledge to facilitate clustering. We summarize such a novel paradigm as externally guided clustering, which provides an innovative perspective on the construction of supervision signals.

- To implement and validate our idea, we propose an externally guided clustering method TAC, which leverages the textual semantics to facilitate image clustering. Experiments demonstrate the superiority of TAC over

eight datasets, including ImageNet-1K. Impressively, in most cases, TAC even outperforms zero-shot CLIP in the absence of class name priors.

- The significance of TAC is two-fold. On the one hand, it proves the effectiveness and superiority of the proposed externally guided clustering paradigm. On the other hand, it suggests the presence of more simple but effective strategies for mining the zero-shot learning ability inherent in VLP models.

2. Related Work

In this section, we review the deep clustering methods, and the zero-shot classification paradigm of VLP models which also utilizes the text modality to perform visual tasks.

2.1. Deep Image Clustering

In addition to effective clustering strategies, discriminative features also play an important role in clustering. Benefiting from the powerful feature extraction ability of neural networks, deep clustering methods show their superiority in handling complex and high-dimensional data (Peng et al., 2016; Guo et al., 2017; Ghasedi Dizaji et al., 2017). The pioneers in deep clustering focus on learning clustering-favorable features through optimizing the network with clustering objectives (Yang et al., 2016; Xie et al., 2016; Peng et al., 2018; Hu et al., 2017; Huang et al., 2020; Ji et al., 2019). In recent years, motivated by the success of contrastive learning, a series of contrastive clustering methods achieve substantial performance leaps on image clustering benchmarks (Li et al., 2021b; Shen et al., 2021; Zhong et al., 2021). Instead of clustering images in an end-to-end manner, several works initially learn image embedding through uni-modal pre-training and subsequently mine clusters based on neighborhood consistency (Van et al., 2020; Dang et al., 2021a) or pseudo-labeling (Niu et al., 2022). By disentangling representation learning and clustering, these multi-stage methods enjoy higher flexibility for their easy adaption to superior pre-trained models. A recent study (Huang et al., 2022) demonstrates that the clustering performance could be further improved when equipping clustering models with more advanced representation learning methods (Grill et al., 2020). Very recently, SIC (Cai et al., 2023) attempts to generate image pseudo labels from the textual space.

Though having achieved remarkable progressions, almost all existing deep image clustering methods mine supervision signals internally. However, the internal supervision signals are inherently bounded by the given images. Instead of pursuing internal supervision signals following previous works, we propose a new paradigm that leverages external knowledge to facilitate image clustering. We hope the simple design and engaging performance of TAC could attract more attention to the externally guided clustering paradigm.

2.2. Zero-shot Classification

Recently, more and more efforts have been devoted to multi-modal, especially vision-language pre-training (VLP). By learning the abundant image-text pairs on the Internet, VLP methods (Li et al., 2021a; Wang et al., 2022), have achieved impressive performance in multi-modal representation learning. More importantly, unlike uni-modal pre-trained models that require additional fine-tuning, VLP models could adapt to various tasks such as classification (Radford et al., 2021), segmentation (Zhou et al., 2022), and image captioning (Li et al., 2022) in a zero-shot manner. Here, we briefly introduce the zero-shot image classification paradigm in CLIP (Radford et al., 2021) as an example. Given names of K classes, CLIP first assembles them with prompts like "a photo of [CLASS]", where the [CLASS] token is replaced by the specific class name. Then, CLIP computes the text embeddings $\{w_i\}_{i=1}^{K}$ of the prompted sentences with its pre-trained text encoder. Finally, CLIP treats the embeddings $\{w_i\}_{i=1}^{K}$ as the classifier weight, and predicts the probability of image $\mathbf{v}$ belonging to the i-th class as

$$p(y = i|\mathbf{v}) = \frac{\exp(\mathrm{sim}(v, w_i)/\tau)}{\sum_{j=1}^{K} \exp(\mathrm{sim}(v, w_j)/\tau)}, \qquad (1)$$

where v denotes the image features, $\mathrm{sim}(\cdot, \cdot)$ refers to the cosine similarity, and τ is the learned *softmax* temperature.

Thanks to the consistent form between pre-training and inference, CLIP achieves promising results in zero-shot image classification. However, such a paradigm requires prior knowledge of class names, which is unavailable in clustering. To leverage CLIP for image clustering, the most direct approach is performing k-means (MacQueen et al., 1967) on the image embeddings. Nevertheless, the performance of k-means is limited and the textual semantics are underutilized. In this work, we explore a more advanced paradigm for image clustering, by taking full advantage of both the pre-trained image and text encoders. Intriguingly, experiments demonstrate that even in the absence of class name priors, the proposed TAC outperforms zero-shot CLIP in most cases. We hope this work could bring some insights into the paradigm design of leveraging VLP models for downstream classification and clustering.

3. Method

In this section, we present TAC, a simple yet effective externally guided clustering method illustrated in Fig. 3. In brief, we first propose a text counterpart construction strategy to exploit the text modality in Sec. 3.1. Then, we propose a cross-modal mutual distillation strategy to collaborate the text and image modalities in Sec. 3.2.
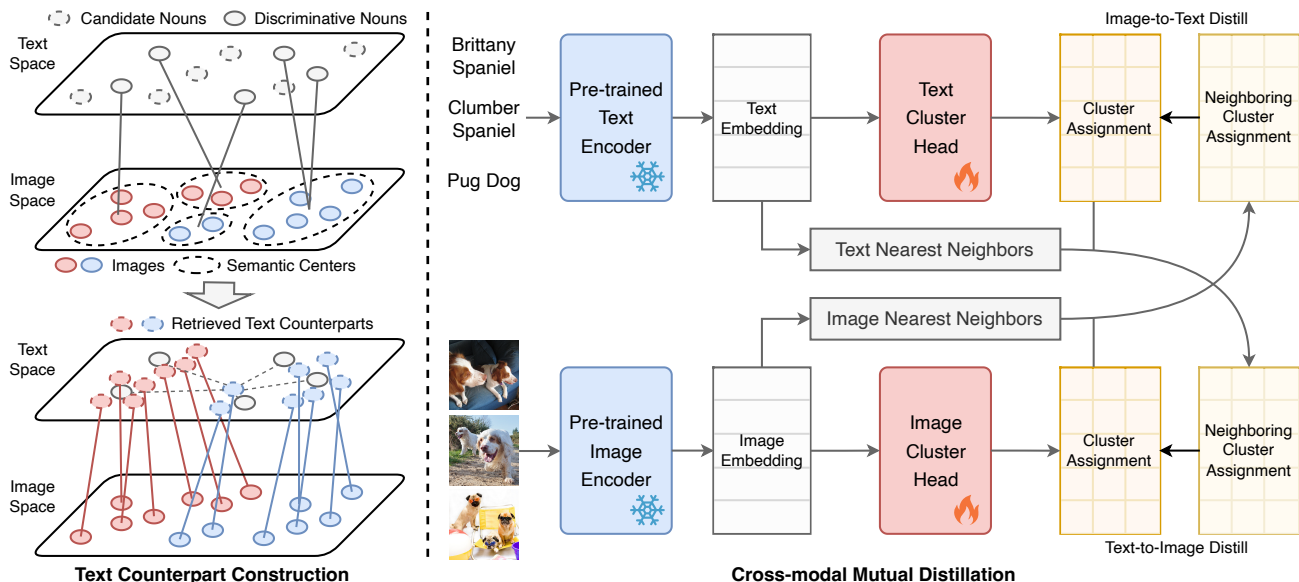
Figure 3. Overview of the proposed TAC. **(Left)** TAC first classifies all nouns from WordNet to image semantic centers, and selects the most discriminative nouns to construct the text space. After that, TAC retrieves nouns for each image to compute its counterpart in the text space. By concatenating the image and retrieved text, we arrive at an extremely simple baseline without any additional training. **(Right)** To better collaborate the text and image modalities, TAC trains cluster heads by mutually distilling the neighborhood information. In brief, TAC encourages images to have consistent cluster assignments with the nearest neighbors of their counterparts in the text embedding space, and vice versa. Such a cross-modal mutual distillation strategy further boosts the clustering performance of TAC.

3.1. Text Counterpart Construction

The textual semantics are naturally favored in discriminative tasks such as classification and clustering. Ideally, clustering could be easily achieved if images have highly distinguishable counterparts in the text modality. To this end, in the absence of class name priors, we propose to select a subset of nouns from WordNet (Miller, 1995) to compose the text space, which is expected to exhibit the following two merits, namely, *i)* precisely covering the image semantics; and *ii)* highly distinguishable between images of different semantics.

The image semantics of different granularities could be captured by k-means with various choices of k. A small value of k corresponds to coarse-grained semantics, which might not be precise enough to cover the semantics of images at cluster boundaries. Oppositely, a large value of k produces fine-grained semantics, which might fail to distinguish images from different classes. To find image semantics of appropriate granularity, we estimate $k = N/300$ given N images, hypothesizing a cluster of $\tilde{N} = 300$ images is compact enough to be described by the same set of nouns. Experiments in Section 4.4.1 show that our TAC is robust across a reasonable range of $\tilde{N}$. With the estimated value of k, we apply k-means on image embeddings to compute the

image semantic centers by

$$s_l = \sum_{i=1}^{N} \mathbb{1}_{v_i \in l} \, v_i, l \in [1, k], \tag{2}$$

where $\mathbb{1}_{v_i \in l}$ is the indicator which equals one iff image v_i belongs to the l-th cluster.

Next, we aim to find discriminative nouns to describe each semantic center. Here, motivated by the zero-shot classification paradigm of CLIP, we reversely classify all nouns from WordNet into k image semantic centers. Specifically, the probability of the i-th noun belonging to the l-th image semantic center is

$$p(y = l | \mathbf{t_i}) = \frac{\exp(\text{sim}(t_i, s_l))}{\sum_{j=1}^{k} \exp(\text{sim}(t_i, s_j))}, \tag{3}$$

where $\mathbf{t_i}$ denoted the i-th noun prompted like CLIP, and t_i is the feature extracted by the text encoder. To identify highly representative and distinguishable nouns, we select the top γ confident nouns for each image semantic center. Formally, the i-th noun would be select for the l-th center if

$$p(y = k | \mathbf{t_i}) \geq \bar{p}(y = k), \tag{4}$$

$$\bar{p}(y = k) = \text{sort}\{p(y = k | \mathbf{t_i}) \,|\, \text{argmax}\, p(y | \mathbf{t_i}) = k\}[\gamma],$$

where $\bar{p}(y = k)$ corresponds to the γ-th largest confidence of nouns belonging to the l-th center. In practice, we fix $\gamma = 5$ on all datasets.

The selected nouns compose the text space catering to the input images. Then, we retrieve nouns for each image to compute its counterpart in the text modality. To be specific, let $\{\bar{\mathbf{t_i}}\}_{i=1}^M$ be the set of M selected nouns with $\{\bar{t}_i\}_{i=1}^M$ being their text embeddings, we compute the text counterpart $\tilde{t}_i$ for image v_i as

$$\tilde{t}_i = \sum_{j=1}^M p(\bar{t}_j|v_i)\bar{t}_j, \tag{5}$$

$$p(\bar{t}_j|v_i) = \frac{\exp(\text{sim}(v_i,\bar{t}_j)/\tilde{\tau})}{\sum_{k=1}^M \exp(\text{sim}(v_i,\bar{t}_k)/\tilde{\tau})}, \tag{6}$$

where $\tilde{\tau} = 0.005$ controls the softness of retrieval. The design of soft retrieval is to prevent the text counterparts of different images from collapsing to the same point. After the text counterpart construction, we arrive at an extremely simple baseline by applying k-means on the concatenated features $[\tilde{t}_i, v_i]_{i=1}^N$. Notably, such an implementation requires no additional training or modifications on CLIP, but it could significantly improve the clustering performance compared with directly applying k-means on the image embeddings (see Section 4.2).

3.2. Cross-modal Mutual Distillation

Though concatenating text counterparts and image embeddings improves the k-means performance, it is suboptimal for collaborating the two modalities. To better utilize multimodal features, we propose the cross-modal mutual distillation strategy. Specifically, let $\mathcal{N}(v_i)$ be a random nearest neighbor of v_i, we introduce a cluster head $f : v \to p \in \mathcal{R}^K$ to predict the soft cluster assignments for images v_i and $\mathcal{N}(v_i)$, where K is the target cluster number. Formally, we denote the soft cluster assignments for n images and their neighbors as

$$P = \begin{bmatrix} p_1 \\ \cdots \\ p_n \end{bmatrix} \text{ and } P^{\mathcal{N}} = \begin{bmatrix} p_1^{\mathcal{N}} \\ \cdots \\ p_n^{\mathcal{N}} \end{bmatrix}. \tag{7}$$

Likewise, we introduce another cluster head $g : \tilde{t}_i \to q_i \in \mathcal{R}^K$ to predict the soft cluster assignments for text counterpart $\tilde{t}_i$ and its random nearest neighbor $\mathcal{N}(\tilde{t}_i)$, resulting in the cluster assignment matrices

$$Q = \begin{bmatrix} q_1 \\ \cdots \\ q_n \end{bmatrix} \text{ and } Q^{\mathcal{N}} = \begin{bmatrix} q_1^{\mathcal{N}} \\ \cdots \\ q_n^{\mathcal{N}} \end{bmatrix}. \tag{8}$$

Let $\hat{p}_i, \hat{p}_i^{\mathcal{N}}, \hat{q}_i, \hat{q}_i^{\mathcal{N}}$ be the i-th column of assignment matrices $P, P^{\mathcal{N}}, Q, Q^{\mathcal{N}}$, the cross-modal mutual distillation loss is

defined as follows, namely,

$$L_{\text{Dis}} = \sum_{i=1}^K L_i^{v \to t} + L_i^{t \to v}, \tag{9}$$

$$L_i^{v \to t} = -\log \frac{e^{(\text{sim}(\hat{q}_i,\hat{p}_i^{\mathcal{N}})/\hat{\tau})}}{\sum_k e^{(\text{sim}(\hat{q}_i,\hat{p}_k^{\mathcal{N}})/\hat{\tau})} + \sum_{k\neq i} e^{(\text{sim}(\hat{q}_i,\hat{q}_k)/\hat{\tau})}}, \tag{10}$$

$$L_i^{t \to v} = -\log \frac{e^{(\text{sim}(\hat{p}_i,\hat{q}_i^{\mathcal{N}})/\hat{\tau})}}{\sum_k e^{(\text{sim}(\hat{p}_i,\hat{q}_k^{\mathcal{N}})/\hat{\tau})} + \sum_{k\neq i} e^{(\text{sim}(\hat{p}_i,\hat{p}_k)/\hat{\tau})}}, \tag{11}$$

where $\hat{\tau}$ is the *softmax* temperature parameter. The distillation loss L_{Dis} has two effects. On the one hand, it minimizes the between-cluster similarity, leading to more discriminative clusters. On the other hand, it encourages consistent clustering assignments between each image and the neighbors of its text counterpart, and vice versa. In other words, it mutually distills the neighborhood information between the text and image modalities, bootstrapping the clustering performance in both. In practice, we set the number of nearest neighbors $\hat{N} = 50$ on all datasets. Note that the neighbors are only computed once on all samples before training.

Next, we introduce two regularization terms to stabilize the training. First, to encourage the model to produce more confident cluster assignments, we introduce the following confidence loss, namely,

$$L_{\text{Con}} = -\log \sum_{i=1}^n p_i^\top q_i, \tag{12}$$

which would be minimized when both p_i and q_i become one-hot. Second, to prevent all samples from collapsing into only a few clusters, we adopt the balance loss, *i.e.*,

$$L_{\text{Bal}} = -\sum_{i=1}^K \left(\bar{p}_i \log \bar{p}_i + \bar{q}_i \log \bar{q}_i \right), \tag{13}$$

$$\bar{p} = \frac{1}{n} \sum_{i=1}^n p_i \in \mathcal{R}^K, \bar{q} = \frac{1}{n} \sum_{i=1}^n q_i \in \mathcal{R}^K, \tag{14}$$

where $\bar{p}$ and $\bar{q}$ correspond to the cluster assignment distribution in the image and text modality, respectively.

Finally, we arrive at the overall objective function of TAC, which lies in the form of

$$L_{\text{TAC}} = L_{\text{Dis}} + L_{\text{Con}} - \alpha \cdot L_{\text{Bal}}, \tag{15}$$

where $\alpha = 5$ is the weight parameter.

4. Experiments

In this section, we evaluate the proposed TAC on five widely used and three more challenging image clustering datasets.

A series of quantitative and qualitative comparisons, ablation studies, and hyper-parameter analyses are carried out to investigate the effectiveness and robustness of the method.

4.1. Experimental Setup

We first introduce the datasets and metrics used for evaluation, and then provide the implementation details of TAC.

4.1.1. DATASETS

To evaluate the performance of our TAC, we first apply it to five widely-used image clustering datasets including STL-10 (Coates et al., 2011), CIFAR-10 (Krizhevsky & Hinton, 2009), CIFAR-20 (Krizhevsky & Hinton, 2009), ImageNet-10 (Chang et al., 2017b), and ImageNet-Dogs (Chang et al., 2017b). With the rapid development of pre-training and clustering methods, we find clustering on relatively simple datasets such as STL-10 and CIFAR-10 is no longer challenging. Thus, we further evaluate the proposed TAC on three more complex datasets with larger cluster numbers, including DTD (Cimpoi et al., 2014), UCF-101 (Soomro et al., 2012), and ImageNet-1K (Deng et al., 2009). Following recent deep clustering works (Van et al., 2020; Dang et al., 2021a), we train and evaluate TAC on the train and test splits, respectively. The brief information of all datasets used in our evaluation is summarized in Table 1.

Table 1. A summary of datasets used for evaluation.

Dataset	Training Split	Test Split	# Training	# Test	# Classes
STL-10	Train	Test	5,000	8,000	10
CIFAR-10	Train	Test	50,000	10,000	10
CIFAR-20	Train	Test	50,000	10,000	20
ImageNet-10	Train	Val	13,000	500	10
ImageNet-Dogs	Train	Val	19,500	750	15
DTD	Train+Val	Test	3,760	1,880	47
UCF-101	Train	Val	9,537	3.783	101
ImageNet-1K	Train	Val	1,281,167	50,000	1,000

4.1.2. EVALUATION METRICS

We adopt three widely-used metrics including Normalized Mutual Information (NMI), Accuracy (ACC), and Adjusted Rand Index (ARI) to evaluate the clustering performance. Higher values of these metrics indicate better results.

4.1.3. IMPLEMENTATION DETAILS

Following previous works (Cai et al., 2023), we adopt the pre-trained CLIP model with ViT-B/32 (Dosovitskiy et al., 2020) and Transformer (Vaswani et al., 2017) as image and text backbones, respectively. For nouns from WordNet (Miller, 1995), we assemble them with prompts like "A photo of [CLASS]" before feeding them into the Transformer. For datasets with an average cluster size less than 300, we empirically set k in k-means thrice as the target cluster number K. The two cluster heads f and g are two-layer MLPs of dimension 512-512-K. We train f and g by the Adam optimizer with an initial learning rate of $1e-3$ for 20 epochs, with a batch size of 512. We fix $\tau = 5e-3$, $\hat{\tau} = 0.5$, and $\alpha = 5.0$ in all the experiments. The only exception is that on UCF-101 and ImageNet-1K, we change $\hat{\tau}$ to 5.0, batch size to 8192, and training epochs to 100, catering to the large cluster number. All experiments are conducted on a single Nvidia RTX 3090 GPU. In our experiments, it takes only one minute to train TAC on the CIFAR-10 dataset.

4.2. Main Results

Here we compare TAC with state-of-the-art baselines on five classic and three more challenging image clustering datasets, followed by feature visualizations to show the superiority of the proposed TAC.

4.2.1. PERFORMANCE ON CLASSIC DATASETS

We first evaluate the proposed TAC on five widely-used image clustering datasets, compared with 15 deep clustering baselines. While early baselines adopt ResNet-34(18) as the backbone, here we mainly focus on comparisons with zero-shot CLIP and CLIP-based methods. As shown in Table 2, by simply retrieving a text counterpart for each image, the proposed TAC successfully mines "free" semantic information from the text encoder. Without any additional training, TAC (no train) substantially improves the k-means clustering performance, especially on more complex datasets. For example, it achieves 14.4% and 43.5% ARI improvements on CIFAR-20 and ImageNet-Dogs, respectively. When further enhanced with the proposed cross-modal mutual distillation strategy, TAC achieves state-of-the-art clustering performance, even surpassing zero-shot CLIP on all five datasets. Such compelling results demonstrate that beyond the current zero-shot classification paradigm, alternative simple but more effective strategies exist for mining the VLP model's ability in image classification and clustering.

4.2.2. PERFORMANCE ON CHALLENGING DATASETS

The clustering results of TAC and baseline methods on more challenging datasets are provided in Table 3. Firstly, we observe TAC without additional training could consistently boost the k-means performance, which achieves a 10% improvement in clustering accuracy on ImageNet-1K. Secondly, although zero-shot CLIP yields slightly better performance on ImageNet-1K given its substantial prior knowledge of 1K class names, TAC still achieves superior performance on DTD and UCF-101 without the class name prior. Such a result verifies the effectiveness of the proposed text counterpart construction strategy, as well as our observation that manually annotated class names are not always the best semantic description.

Table 2. Clustering performance on five widely-used image clustering datasets. The best and second best results are denoted in **bold** and underline, respectively.

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 et al., 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 et al., 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., 2017a)	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 et al., 2019)	49.6	59.6	39.7	51.3	61.7	41.1	22.5	25.7	11.7	–	–	–	–	–	–	–
PICA (Huang et al., 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., 2021b)	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 et al., 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 et al., 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., 2021a)	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 et al., 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)	<u>95.3</u>	<u>98.1</u>	<u>95.9</u>	**84.7**	**92.6**	**84.4**	59.3	<u>58.3</u>	<u>43.9</u>	97.0	98.2	96.1	69.0	69.7	55.8	<u>79.9</u>
CLIP (k-means)	91.7	94.3	89.1	70.3	74.2	61.6	49.9	45.5	28.3	96.9	98.2	96.1	39.8	38.1	20.1	66.3
TAC (no train)	92.3	94.5	89.5	80.8	90.1	79.8	<u>60.7</u>	55.8	42.7	<u>97.5</u>	<u>98.6</u>	<u>97.0</u>	<u>75.1</u>	<u>75.1</u>	<u>63.6</u>	79.5
TAC	**95.5**	**98.2**	**96.1**	<u>83.3</u>	<u>91.9</u>	<u>83.1</u>	**61.1**	**60.7**	**44.8**	**98.5**	**99.2**	**98.3**	**80.6**	**83.0**	**72.2**	**83.2**
CLIP (zero-shot)	93.9	97.1	93.7	80.7	90.0	79.3	55.3	58.3	39.8	95.8	97.6	94.9	73.5	72.8	58.2	78.7

Table 3. Clustering performance on three more challenging image clustering datasets. The best and second best results are denoted in **bold** and underline, respectively.

Dataset	DTD			UCF-101			ImageNet-1K			AVG
Metrics	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	
CLIP (k-means) (Radford et al., 2021)	57.3	42.6	27.4	79.5	58.2	47.6	72.3	38.9	27.1	50.1
SCAN (Van et al., 2020)	59.4	<u>46.4</u>	<u>31.7</u>	79.7	61.1	53.1	74.7	44.7	32.4	53.7
SIC (Cai et al., 2023)	59.6	45.9	30.5	81.0	<u>61.9</u>	<u>53.6</u>	77.2	47.0	34.3	54.6
TAC (no train)	<u>60.1</u>	45.9	29.0	<u>81.6</u>	61.3	52.4	<u>77.8</u>	<u>48.9</u>	<u>36.4</u>	<u>54.8</u>
TAC	**62.1**	**50.1**	**34.4**	**82.3**	**68.7**	**60.1**	**79.9**	**58.2**	**43.5**	**59.9**
CLIP (zero-shot) (Radford et al., 2021)	56.5	43.1	26.9	79.9	63.4	50.2	81.0	63.6	45.4	56.7

4.2.3. VISUALIZATION

To provide an intuitive understanding of the clustering results, we visualize the features obtained at four different steps of TAC in Fig. 4. The clustering performance by applying k-means on the features is annotated at the top. Fig. 4(a) shows the image features extracted by the pretrained CLIP image encoder. As can be seen, images of different breeds of dogs are mixed, leading to a poor clustering ARI of 23.4%. By selecting and retrieving discriminative nouns, visually similar samples could be better distinguished in the text modality as shown in Fig. 4(b). By simply concatenating images and retrieved text counterparts, TAC significantly improves the feature discriminability and k-means performance without any additional training. Finally, when equipped with the proposed cross-modal mutual distillation strategy, TAC could better collaborate the image and text modalities, leading to the best within-clustering compactness and between-cluster scatterness.

4.3. Ablation Study

In this section, we conduct ablation studies on the three loss terms and the direction of the cross-modal distillation.

4.3.1. LOSS TERMS

To understand the efficacy of the three loss terms L_{Dis}, L_{Con}, and L_{Bal} in Eq. 9, 12, and 13, we evaluate the performance of TAC with different loss combinations. According to the results in Table 4, one could see that: *i)* the balance loss L_{Bal} could prevent cluster collapsing. Without L_{Bal}, TAC assigns most images to only a few clusters, leading to poor clustering performance on both datasets; *ii)* the confidence loss L_{Con} is necessary for datasets with large cluster numbers. The reason is that the cluster assignments would be less confident when the cluster number is large. In this case, the regularization efficacy of L_{Bal} would be alleviated, which explains the performance degradation on UCF-101; and *iii)* L_{Dis} could effectively distill the neighborhood information between the text and image modalities, leading

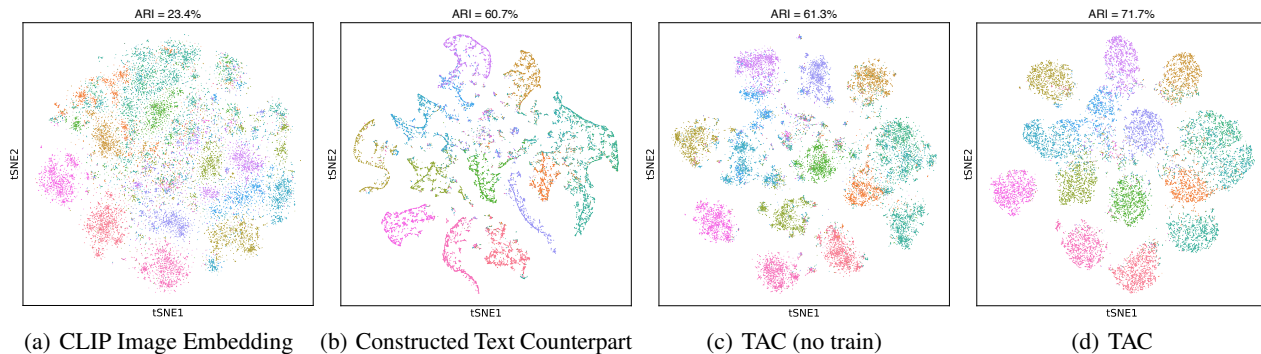(a) CLIP Image Embedding   (b) Constructed Text Counterpart   (c) TAC (no train)   (d) TAC

Figure 4. Visualization of features extracted by different methods on the ImageNet-Dogs training set, with the corresponding k-means clustering ARI annotated on the top. a) image embedding directly obtained from the CLIP image encoder; b) text counterparts constructed by TAC; c) concatenation of images and text counterparts; d) representation learned by TAC through cross-modal mutual distillation.

to the best clustering performance.

Table 4. The performance of TAC with different combinations of the loss terms.

L_{Dis}	L_{Con}	L_{Bal}	ImageNet-Dogs			UCF-101		
			NMI	ACC	ARI	NMI	ACC	ARI
✓			71.4	69.5	38.1	69.3	7.6	13.6
	✓		57.2	14.3	24.3	52.1	3.4	8.6
		✓	15.1	19.3	4.1	43.5	16.2	5.7
✓	✓		72.5	57.0	45.3	55.6	3.6	9.9
✓		✓	**80.6**	**83.5**	**72.3**	70.5	45.1	34.5
	✓	✓	78.2	81.8	69.6	<u>81.6</u>	<u>67.3</u>	<u>59.1</u>
✓	✓	✓	**80.6**	<u>83.0</u>	<u>72.2</u>	**82.3**	**68.7**	**60.1**

4.3.2. DISTILLATION DIRECTION

Recall that the cross-modal distillation strategy mutually distills the neighborhood information from one modality to another. To better understand the effectiveness of mutual distillation, we evaluate the performance of TAC with different directions of the distillation in Table 5. As can be seen, text-to-image distillation gives inferior performance compared with bi-directional distillation, probably due to the less exploration of image neighborhood information. Moreover, in the one-directional scenarios, text-to-image outperforms image-to-text distillation, which proves that the textual semantics are more favorable for clustering.

Table 5. The performance of TAC with different distillation directions. †: Use the text head for clustering.

Direction	ImageNet-Dogs			UCF-101		
	NMI	ACC	ARI	NMI	ACC	ARI
Image $\rightarrow$ Text[†]	76.5	79.3	67.1	78.5	64.1	53.7
Text $\rightarrow$ Image	78.8	82.1	69.4	81.1	65.8	57.5
Image $\leftrightarrow$ Text	**80.6**	**83.0**	**72.2**	**82.3**	**68.7**	**60.1**

4.4. Parameter Analyses

To evaluate the robustness of TAC, we evaluate it under various choices of the expected compact cluster size $\tilde{N}$, the number of discriminative nouns for each image semantic center γ, and the number of nearest neighbors $\hat{N}$. The results are shown in Fig. 5.

4.4.1. EXPECTED COMPACT CLUSTER SIZE $\tilde{N}$

To see how the granularity of image semantics influences the final clustering performance, we test various choices of $\tilde{N}$ in Fig. 5(a). As shown, TAC is stable across a reasonable range of $\tilde{N}$. However, since UCF-101 has an average cluster size much less than the default $\tilde{N} = 300$, it encounters a performance drop on large cluster sizes due to overly coarse-grained semantics. Conversely, when the cluster size is overly small, the excessively fine-grained semantics leads to performance degradation on ImageNet-Dogs.

4.4.2. NUMBER OF DISCRIMINATIVE NOUNS γ

To construct the text space, we classify all nouns into the image semantic centers and select the top γ nouns of each center for retrieval. Here, we try various choices of γ in Fig. 5(b). As can be seen, a solitary noun is insufficient to cover the semantics of each image center. Conversely, an excessive number of nouns would falsely enrich the semantics, leading to inferior performance. Overall, TAC is stable across a typical range of discriminative noun number γ.

4.4.3. NUMBER OF NEAREST NEIGHBORS $\hat{N}$

To collaborate the text and image modalities, TAC mutually distills their neighborhood information. Here, we evaluate TAC with different numbers of nearest neighbors $\bar{N}$ in Fig. 5(c). The results demonstrate that TAC is robust to diverse numbers of $\bar{N}$. Though a smaller choice of $\bar{N}$ leads to slight improvements on ImageNet-Dogs, we find the default
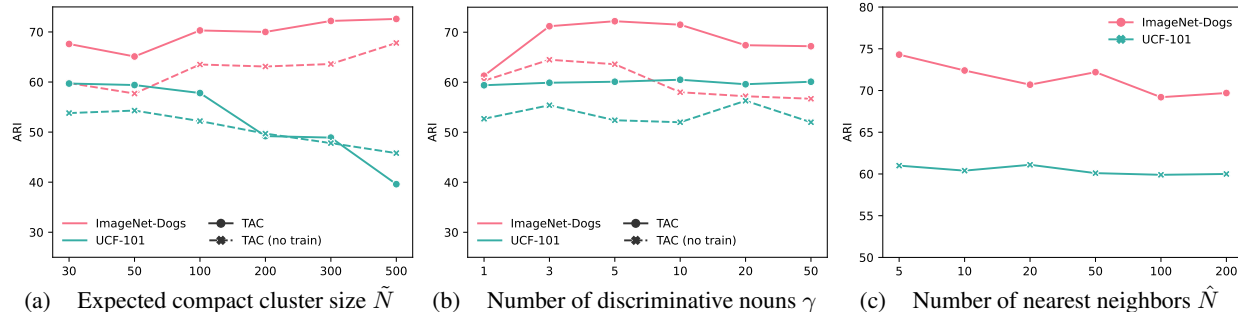
Figure 5. Analyses on six tunable hyper-parameters in the proposed TAC. The first three hyper-parameters influence both TAC with and without training. The last three hyper-parameters only influence the cross-modal mutual distillation process of TAC.

$\bar{N} = 50$ achieves stable results across different datasets.

5. Conclusion

In this paper, instead of focusing on exhaustive internal supervision signal construction, we innovatively propose leveraging the rich external knowledge, which has been regrettably overlooked before, to facilitate clustering. As a specific implementation, our TAC achieves state-of-the-art image clustering performance by leveraging textual semantics, demonstrating the effectiveness and promising prospect of the proposed externally guided clustering paradigm. In the future, the following directions could be worth exploring. On the one hand, in addition to the modalities this work focuses on, the external knowledge widely exists in different sources, domains, models, etc. For example, one could utilize the pre-trained object detection or semantic segmentation models to locate the semantic object for boosting image clustering. On the other hand, instead of focusing on image clustering, it is worth exploring the external knowledge for clustering other forms of data, such as text and point cloud. The challenges of the proposed externally guided clustering paradigm lie in *i)*, choosing the appropriate external knowledge, and *ii)*, effectively integrating the external knowledge to improve clustering. In practice, the selection and utilization of external knowledge should depend on the characteristics and prior knowledge about the data and task. Overall, we hope this work could serve as a catalyst, motivating more future studies on externally guided clustering, which is believable to be a promising direction for both methodology improvement and real-world application.

Acknowledgements

Impact Statement

This work proposes a new deep clustering paradigm by leveraging external knowledge. As a fundamental problem in machine learning, clustering has a wide range of applications, such as anomaly detection, person re-identification, community detection, etc. The proposed method is evaluated on public image datasets that are not at risk. However, just like any learning method, the performance of our method depends on data bias and cannot be guaranteed in more complex real-world applications. In this sense, it might bring some disturbances in decision-making and thus should be carefully used, especially in areas such as health care, autonomous vehicles, etc. Moreover, the proposed method requires manually setting the target cluster number. In real-world applications, one may resort to other cluster number estimation methods in the lack of cluster number prior.

References

Cai, D., He, X., Wang, X., Bao, H., and Han, J. Locality preserving nonnegative matrix factorization. In *International Joint Conference on Artificial Intelligence*, volume 9, pp. 1010–1015, 2009.

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

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

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

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 215–223, 2011.

Dang, Z., Deng, C., Yang, X., Wei, K., and Huang, H. Nearest neighbor matching for deep clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13693–13702, 2021a.

Dang, Z., Deng, C., Yang, X., Wei, K., and Huang, H. Nearest neighbor matching for deep clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13693–13702, 2021b.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Ieee, 2009.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Elhamifar, E. and Vidal, R. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11): 2765–2781, 2013.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pp. 226–231, 1996.

Ghasedi Dizaji, K., Herandi, A., Deng, C., Cai, W., and Huang, H. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In *International Conference on Computer Vision*, pp. 5736–5745, 2017.

Gowda, K. C. and Krishna, G. Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern Recognition*, 10(2):105–112, 1978.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

Guo, X., Liu, X., Zhu, E., and Yin, J. Deep clustering with convolutional autoencoders. In *International Conference on Neural Information Processing*, pp. 373–382. Springer, 2017.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Hu, W., Miyato, T., Tokui, S., Matsumoto, E., and Sugiyama, M. Learning discrete representations via information maximizing self-augmented training. In *International Conference on Machine Learning*, pp. 1558–1567. PMLR, 2017.

Huang, J., Gong, S., and Zhu, X. Deep semantic clustering by partition confidence maximisation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020.

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

Ji, X., Henriques, J. F., and Vedaldi, A. Invariant information clustering for unsupervised image classification and segmentation. In *International Conference on Computer Vision*, pp. 9865–9874, 2019.

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

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., and Wang, H. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Annual Meeting of the Association for Computational Linguistics*, pp. 2592–2607, 2021a.

Li, X., Zhang, R., Wang, Q., and Zhang, H. Autoencoder constrained clustering with adaptive neighbors. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–7, 2020.

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

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., and Ma, Y. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2012.

Liu, W., Shen, X., and Tsang, I. Sparse embedded k-means clustering. In *Advances in Neural Information Processing Systems*, pp. 3319–3327, 2017.

MacQueen, J. et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pp. 281–297. Oakland, CA, USA, 1967.

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

Nie, F., Zeng, Z., Tsang, I. W., Xu, D., and Zhang, C. Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering. *IEEE Transactions on Neural Networks*, 22(11):1796–1808, 2011.

Nie, F., Wang, X., Jordan, M. I., and Huang, H. The constrained laplacian rank algorithm for graph-based clustering. In *AAAI Conference on Artificial Intelligence*, pp. 1969–1976. Citeseer, 2016.

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

Peng, X., Xiao, S., Feng, J., Yau, W.-Y., and Yi, Z. Deep subspace clustering with sparsity prior. In *International Joint Conference on Artificial Intelligence*, pp. 1925–1931, 2016.

Peng, X., Feng, J., Xiao, S., Yau, W.-Y., Zhou, J. T., and Yang, S. Structured autoencoders for subspace clustering. *IEEE Transactions on Image Processing*, 27(10):5076–5086, 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

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

Soomro, K., Zamir, A. R., and Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Tao, Y., Takagi, K., and Nakata, K. Clustering-friendly representation learning via instance discrimination and feature decorrelation. In *International Conference on Learning Representations*, 2020.

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

Van, G. W., Vandenhende, S., Georgoulis, S., Proesmans, M., and Van Gool, L. Scan: Learning to classify images without labels. In *European Conference on Computer Vision*, pp. 268–285. Springer, 2020.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

Wang, Z., Li, Z., Wang, R., Nie, F., and Li, X. Large graph clustering with simultaneous spectral embedding and discretization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

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

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

Yang, J., Parikh, D., and Batra, D. Joint unsupervised learning of deep representations and image clusters. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5147–5156, 2016.

Zelnik-Manor, L. and Perona, P. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems*, pp. 1601–1608, 2005.

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

Zhou, C., Loy, C. C., and Dai, B. Extract free dense labels from clip. In *European Conference on Computer Vision*, pp. 696–712. Springer, 2022.

A. Variants of text counterpart construction.

Table 6. Clustering performance of TAC using different clustering methods for text counterpart construction. (AC: agglomerative clustering, SC: spectral clustering, r: resolution of Louvain clustering, None: using all nouns from WordNet)

Method	Semantic Space	ImageNet-Dogs			UCF-101		
		NMI	ACC	ARI	NMI	ACC	ARI
TAC (no train)	k-means	75.1	75.1	63.6	81.6	61.3	52.4
	AC	73.4	72.0	61.2	81.9	63.7	**54.8**
	SC	**77.4**	75.2	65.9	82.2	**65.5**	54.7
	DBSCAN	68.8	64.3	51.0	81.1	61.8	52.3
	Louvain (r=1)	77.0	75.1	65.0	78.6	58.3	47.9
	Louvain (r=5)	77.1	**78.3**	**66.9**	81.3	61.7	54.0
	Louvain (r=10)	75.7	72.9	62.7	80.9	60.8	52.5
	None	70.3	68.7	53.6	81.3	63.2	52.8
TAC	k-means	**80.6**	83.0	72.2	82.3	68.7	60.1
	AC	78.4	81.7	69.5	**82.4**	**69.3**	60.2
	SC	79.2	83.1	70.8	82.3	69.1	**60.4**
	DBSCAN	75.5	80.4	65.5	80.6	66.2	56.4
	Louvain (r=1)	78.5	83.6	70.0	78.3	61.7	53.0
	Louvain (r=5)	79.8	**85.6**	**72.6**	81.7	68.3	59.1
	Louvain (r=10)	79.6	85.1	71.9	82.1	68.1	59.4
	None	75.7	78.7	66.0	81.2	67.0	58.3

Recall that to select representation nouns for text counterpart construction, we first classify all nouns from WordNet to image semantic centers found by applying k-means on image embeddings. Here, we investigate the robustness and necessity of the noun selection step. Specifically, we adopt three other classic clustering methods to compute semantic centers, including agglomerative clustering (AC), spectral clustering (SC), and DBSCAN. For AC and SC, we set the target cluster number to the same as k-means. For DBSCAN, we tune the density parameter until it produces the same number of clusters. As shown in Table 6, the training-free TAC achieves better performance with SC, while the performance is similar among k-means, AC, and SC when further boosted with cross-modal mutual distillation. The performance degradation on DBSCAN is probably due to the poor quality of image embeddings. In practice, we find DBSCAN tends to treat a portion of samples as outliers, and thus it cannot precisely cover the image semantics, leading to suboptimal performance. Moreover, we test the Louvain clustering algorithm that could estimate the cluster number given the cluster resolution. One could see that Louvain clustering gives promising results with an appropriate choice of resolution. Nevertheless, almost all cluster number estimation methods require manually setting a granularity parameter like the resolution here in Louvain. Such a process is similar to our simple estimation based on the sample size. To investigate the necessity to filter discriminative nouns, we further append a baseline by retrieving text counterparts from all nouns. According to the results, TAC encounters a performance drop on both datasets, but the influence is milder on UCF-101, which could be attributed to the richer image semantics in that dataset. In summary, the results demonstrate the effectiveness of discriminative noun selection, as well as the robustness of TAC against different clustering methods used for text counterpart construction.

B. The textual semantic space constructed by TAC.

To provide an intuitive understanding of the textual space constructed by TAC, we provide the discriminative nouns selected for all datasets. Due to the space limitation, only the thirty most discriminative nouns are shown. As can be seen, for object clustering datasets including STL, CIFAR, and ImageNet, the selected discriminative nouns directly match the name of objects. The results are more intriguing on the DTD and UCF-101 datasets with textures and actions as the clustering criterion, respectively. For the DTD dataset, some nouns directly match the adjectives that describe textures. For example, "Belgian waffle" matches the "waffled" texture, and "honeycomb" matches the "honeycombed" texture. For other selected nouns that do not have a direct matching, they turn out to have close relationships with those textures. For example, the nouns "garden lettuce", "peony", and "Peruvian lily" correspond to the "frilly" texture, "grevy's zebra" corresponds to the "striped" texture, and "chessboard" corresponds to the "chequered" texture. In other words, these nouns describe or reflect the texture and can thus benefit the discrimination between images of different textures. For the UCF-101 dataset, most selected nouns correspond to the object that actions interact with. For example, the "snooker table" in "billiard hall" is used

for "Billiards" and the "typewriter keyboard" is used for "Typing". There also exist some gerundial nouns that directly refer to the actions such as "cliff diving" and "touch typing". The close connection between the nouns and actions explains the performance improvement in image clustering. To summarize, external nouns from WordNet would be closely related to the semantics in images, either directly or indirectly. As a result, these nouns could provide more compact semantics and benefit clustering.

Table 7. Top-30 selected discriminative nouns for image semantic centers from different datasets.

Dataset	Selected Discriminative Nouns
STL-10	floatplane, titi monkey, sand cat, whitetail deer, harness horse, black billed cuckoo, garbage truck, fire truck, Lipizzan, container ship, ocean liner, trucking rig, aerobatics, Angora cat, airline, black and tan terrier, electric automobile
CIFAR-10	spadefoot toad, field sparrow, curassow, Lipizzan, sable antelope, black fronted bush shrike, chameleon tree frog, whitetail deer, fire truck, waterbuck, emu, hartebeest, dressage, containership, pratincole, woodland caribou, hydroplane racing, banana boat, yacht race, yellow breasted chat, pen tailed tree shrew, elk, wagtail, stonefish, trucking rig, clipper ship, Texas horned lizard, stealth bomber, articulated lorry, fire department
CIFAR-20	Iceland poppy, Lepiota procera, oceanic whitetip shark, prairie sunflower, goblet, common dolphin, carabid beetle, sunflower, rosy boa, trolleybus, soda can, school bus, Peromyscus maniculatus, characin fish, banded gecko, Arabian camel, armoured combat vehicle, diesel electric locomotive, African elephant, bunk bed, tandem bicycle, sandbar shark, common wallaby, European spider crab, eastern chimpanzee, navel orange, edmontosaurus, Kodiak bear, lawn mower, tractor
ImageNet-10	crested penguin, snow leopard, Graf Zeppelin, navel orange, Maltese terrier, soccer ball, blimp, candied citrus peel, airship, serval, dirigible, containership, articulated lorry, trailer truck, tractor trailer, sports car, tufted puffin, soccer player, wire haired terrier, roadster, Sealyham terrier, soft coated wheaten terrier, sporting dog, airline, turboprop, flying bomb, wind bell, fruit tree, Antarctic Peninsula, airliner
ImageNet-Dogs	Doberman pinscher, giant schnauzer, Norwegian elkhound, clumber spaniel, chowchow, Shetland sheepdog, Welsh springer spaniel, pug, standard schnauzer, schipperke, basset, beach, merino sheep, Arctic wolf, keeshond, Maltese terrier, standard poodle, snowfall, swimming hole, chromolithography, sleeping partner, chipping sparrow, dog show, Persian cat, Pomeranian, golden retriever, triple jump, meadow jumping mouse, fieldwork, harness race
DTD	butterflyfish, garden lettuce, peony, grevy's zebra, chessboard, Peruvian lily, pothole, Belgian waffle, turban squash, African elephant, anchor rope, chainlink fence, wicker basket, sweetsop, honeycomb, komondor, cytologic smear, rood screen, orb weaving spider, pillow lace, fluorite, proboscis monkey, birch bark, grape leaf begonia, sunset, zebrawood, stockinette stitch, lecanora, houndstooth check, black crappie
UCF-101	sitar, cliff diving, snooker table, blackboard, typewriter keyboard, billiard hall, marching band, darning needle, touch typing, Islamic Army of Aden, piano sonata, violoncellist, table tennis, koto player, contradance, bowling alley, Seattle Slew, tai chi chuan, parade, sumo ring, candlepin bowling, pelican crossing, cymbalist, Armenian Secret Army for the Liberation of Armenia, tenor drum, woodwind instrument, Panjabi, Victory Day, cyclist, Belmont Stakes
ImageNet-1K	Cypripedium fasciculatum, fireboat, swamphen, colobus monkey, plaque, komondor, limpkin, chasuble, European black grouse, nine banded armadillo, sulphur crested cockatoo, ruffed grouse, common stinkhorn, genus Cypripedium, Geastrum coronatum, ptarmigan, red breasted merganser, tobacconist shop, oystercatcher, axolotl, slate colored junco, purple gallinule, black capped chickadee, Tibetan mastiff, redshank, red legged partridge, Polaroid camera, pygmy marmoset, cherimoya, sharp tailed grouse