Semantic-Enhanced Image Clustering

Shaotian Cai*, Liping Qiu*, Xiaojun Chen†, Qin Zhang, Longteng Chen

Shenzhen University, Shenzhen, China
cai.st@foxmail.com, qiuliping2021@email.szu.edu.cn, {xjchen, qinzhang}@szu.edu.cn, chenlt2021@163.com

Abstract

Image clustering is an important and open-challenging task in computer vision. Although many methods have been proposed to solve the image clustering task, they only explore images and uncover clusters according to the image features, thus being unable to distinguish visually similar but semantically different images. In this paper, we propose to investigate the task of image clustering with the help of a visual-language pre-training model. Different from the zero-shot setting, in which the class names are known, we only know the number of clusters in this setting. Therefore, how to map images to a proper semantic space and how to cluster images from both image and semantic spaces are two key problems. To solve the above problems, we propose a novel image clustering method guided by the visual-language pre-training model CLIP, named **Semantic-Enhanced Image Clustering (SIC)**. In this new method, we propose a method to map the given images to a proper semantic space first and efficient methods to generate pseudo-labels according to the relationships between images and semantics. Finally, we propose performing clustering with consistency learning in both image space and semantic space, in a self-supervised learning fashion. The theoretical result of convergence analysis shows that our proposed method can converge at a sublinear speed. Theoretical analysis of expectation risk also shows that we can reduce the expected risk by improving neighborhood consistency, increasing prediction confidence, or reducing neighborhood imbalance. Experimental results on five benchmark datasets clearly show the superiority of our new method.

Introduction

Image classification, which assigns an image to a predefined set of classes, is an important task in computer vision. However, it is costly to obtain labeled data in the age of big data. To liberate us from laborious and trivial data labeling work, image clustering that aims to group images into different clusters without ground-truth semantic labels has become a more and more important task.

The early works in deep image clustering usually combine auto-encoders (AE) or Convolutional Neural Network (CNN) based representation learning with traditional shallow clustering methods (Xie, Girshick, and Farhadi 2016; Yang, Parikh, and Batra 2016; Yang et al. 2017; Tian, Zhou, and Guan 2017; Shaham and Stanton 2018). In recent years, with the rapid development of pre-training models, such as VGG-16 (Simonyan and Zisserman 2014), Resnet (He et al. 2016), ViT (Dosovitskiy et al. 2020), Swin Transformer (Liu et al. 2021), image clustering methods leave the images representation task for pre-training model, and directly map image representations into labels by training a classification model like Multilayer Perceptron (MLP), by maximizing the mutual information between the image and its augmentations (Ji, Vedaldi, and Henriques 2019b; Li et al. 2021; Zhong et al. 2021) or the likelihood of the cluster assignments between the image and its neighbors (Wu et al. 2019; Van Gansbeke et al. 2020; Zhong et al. 2021; Dang et al. 2021). However, since we want to obtain semantically meaningful clusters, it is difficult to solve this problem by only exploring images. Figure 1 shows some examples that are semantically different but visually similar. For example, an image with an airplane may be visually similar to an image with a bird, and an image with a deer may be visually similar to an image with a horse.

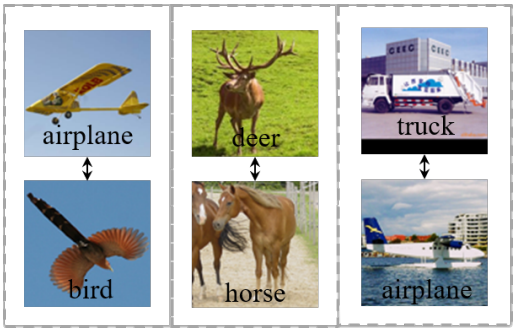


Figure 1: Visually similar but semantically different images on STL10 dataset, where every two images in a column are similar with the image embedding obtained by the CLIP pre-training model and the text in an image is its true label.

Intuitively, we need to access the language model to improve image clustering with semantic information. Some works (Jin et al. 2015; Wang et al. 2020; Yang, Huang, and

*These authors contributed equally.
†Corresponding author.

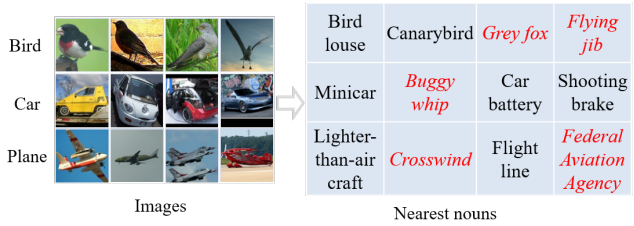


Figure 2: Images and their nearest nouns selected from WordNet (Miller 1995) on STL10, where the image and text embeddings are obtained via CLIP (Radford et al. 2021). The images corresponding to red italic nouns are wrongly mapped.

Howe 2021) try to explore the image-caption pairs to cluster images, but constructing the images with qualified captions is cost-intensive in real applications. Note the great success of visual-language pre-training models such as CLIP (Radford et al. 2021), which is trained on a dataset of 400 million image-text pairs available on the internet to align texts and images in common feature space by capturing the image-text relationships. It has shown surprising results in zero-shot learning tasks. However, we need to know the class names in zero-shot tasks, which hinders some potential applications such as image clustering when the class names are unavailable. This motivates us to utilize the visual-language pre-training model to compensate for the semantic information for better image clustering.

Although visual-language pre-training models such as CLIP can map images and texts into a unified space, Figure 2 shows that simply mapping images to the nearest semantics does not improve the clustering. Therefore, in the task of image clustering with help of a visual-language pre-training model, we need to solve two key problems:

1. *How to map images to a proper semantic space that can improve the clustering?*
2. *How to cluster images from both image and semantic spaces?*

In this paper, as shown in Figure 3, we propose a novel image clustering method guided by the visual-language pre-training model CLIP, named **Semantic-Enhanced Image Clustering (SIC)**. The new method first maps the given images to a proper semantic space, generates pseudo-labels by taking the relationships between images and semantics into consideration, and then performs image clustering with consistency learning in both image space and semantic space. Our main contributions are summarized as follows:

- We propose a method to select proper nouns to construct semantic space, and three methods to map images to semantics in order to generate pseudo-labels.
- The theoretical result on convergence shows that our proposed method can converge at a sublinear speed.
- The theoretical result on expectation risk shows that we can reduce the expected risk of our method by improving neighborhood consistency, increasing prediction confidence, or reducing neighborhood imbalance such that a sample lies in less sample's nearest neighborhoods.
- Experimental results on five benchmark datasets clearly show that SIC is superior to 20 state-of-the-art and zero-shot learning with CLIP.

Related Work

Vision-Language Pre-training Models

Vision-Language Pre-training (VLP) models align multi-modal data in common feature space by different pre-training tasks, which can be categorized into two categories: 1) VisualBert (Li et al. 2019), UNITER (Chen et al. 2020) and DALL-E (Ramesh et al. 2021) use Language-based training strategy, including mask LM (Mask Language Modeling) such as Masked Language/Region Modeling, or autoregressive LM such as image caption and text-grounded image generation. 2) UNIMO (Li et al. 2020b), CLIP (Radford et al. 2021), ALIGN (Jia et al. 2021) utilize cross-modal contrastive learning to align the visual and textual information into a unified semantic space.

The core task of VLP is to model the interactions between images and texts, and there are two types of architectures for this: 1) The single-stream models like ImageBERT (Qi et al. 2020), Unicoder-VL (Li et al. 2020a) concatenate patch-wise or regional visual and textual embeddings and feed them to one encoder. 2) The dual-stream models like ViLBERT (Lu et al. 2019) and CLIP (Radford et al. 2021) obtain visual and textual embeddings with separate encoders.

Since VLP captures the relationships among images and texts (low-level semantics), in this paper, we propose to utilize the visual-language pre-training model to compensate for the semantic information for better image clustering.

Image Clustering

The early works in deep clustering usually simply combined feature learning with shallow clustering. For example, some methods combined the stacked auto-encoders (SAE) with the traditional clustering algorithms such as k-means (Xie, Girshick, and Farhadi 2016; Yang et al. 2017; Tian, Zhou, and Guan 2017), subspace clustering (Ji et al. 2017) and spectral clustering (Shaham and Stanton 2018), or combined the Convolutional Neural Network (CNN) with the hierarchical clustering (Yang, Parikh, and Batra 2016). However, the above methods usually require post-processing to obtain cluster assignments.

Recently, some methods were developed to directly map images into labels with a classification model, by maximizing the mutual information between the labels of the original images and their augmentations (Ji, Vedaldi, and Henriques 2019a; Li et al. 2021; Zhong et al. 2021), or maximizing the likelihood of the cluster assignments between a sample and its nearest neighbors (Zhong et al. 2021; Dang et al. 2021; Chang et al. 2017a; Wu et al. 2019; Van Gansbeke et al. 2020). Some of them further generate pseudo-labels to refine the model (Wu et al. 2019; Van Gansbeke et al. 2020). Furthermore, some methods were proposed to act as add-on modules to revise the classification model via label cleans-

ing and retraining with the refined labels (Gupta et al. 2020; Park et al. 2021).

The pseudo-labels in (Wu et al. 2019; Van Gansbeke et al. 2020) are generated from the clustering results, and thus are doubtful. (Mahon and Lukasiewicz 2021) generates multiple groups of pseudo-labels by training multiple clustering algorithms independently, and sets the common pseudo-labels as high-quality pseudo-labels. However, it is cost-intensive, and the Hungarian algorithm makes it difficult to effectively align multiple groups of pseudo-labels.

In this paper, we propose to generate high-quality pseudo-labels according to the interaction between image and text by utilizing the vision-language model CLIP.

Notation and Problem Definition

In this paper, matrices are written as bold uppercase letters like $\mathbf{A}$. $\mathbf{a}_i$ represents the i-th row of $\mathbf{A}$, a_{ij} represents the i-th row and the j-th column element of $\mathbf{A}$ and $\mathbf{A}^T$ is the transpose of $\mathbf{A}$. $\|\cdot\|_1$ expresses the l_1-norm of a vector. $\|\cdot\|$ donates the module of vector.

Suppose we have an image dataset with n instances sampled i.i.d. from input space $\mathcal{X}$ is denoted as $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$, we can obtain the embeddings of these images as $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ where $\mathbf{u}_i = g(x_i)$ is obtained via the image encoder $g(.)$ of CLIP. To capture the semantic meaning of these images, we introduce a semantic dataset $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}$ that includes m noun phrases from WordNet (Miller 1995) and define a function $h(t_i)$ to obtain the embedding of each noun t_i from CLIP (Radford et al. 2021), by constructing a sentence s_i like "`A photo of a` $\{t_i\}$" and obtain their semantic embeddings as $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\}$ where $\mathbf{v}_i = h(s_i)$ from the text encoder of CLIP. Let c be the number of categories; our goal is to group the images in $\mathcal{D}$ into c clusters with the help of the CLIP model. Let $f(g(\mathcal{D}); \phi) : \mathcal{V} \to \mathbb{R}^c$ denotes the network with parameters ϕ that maps an image x_i with embedding $\mathbf{u}_i$ into soft cluster assignment probability $\mathbf{q}_i$. f is implemented by a multilayer perceptron (MLP). Notably, the image and text encoders in CLIP are kept frozen during the training process, i.e., the parameters in the functions $g(.)$ and $h(.)$ are fixed.

The Proposed Method

In this paper, we propose a novel image clustering method, which is shown in Figure 3. The new method consists of three steps: 1) **Semantic Space Construction** selects meaningful texts to construct semantic space; 2) **Semantic-Enhanced Pseudo-labeling** generates pseudo-labels by taking both image and semantic spaces into consideration; and 3) **Joint Consistency Learning** performs image clustering with the consistency learning in both image and semantic spaces. In the following, we will give the details of the three steps.

Semantic Space Construction

In this step, we aim to construct a semantic space suitable for images by introducing related texts. In an image clustering task, we need to cluster images by their object category attributes, and the set of object category names is usually a subset of the commonly used nouns in the English language. For example, in CIFAR10, the class names are 10 commonly used English nouns ("airplane", "automobile", "bird", etc.). Therefore, we take the entire list of nouns in the WordNet dataset (Miller 1995) to form a semantic dataset $\mathcal{W}$ which contains more than 82, 000 nouns. Since an image dataset usually covers only a small set of categories, we propose a two-step method to select most related nouns from $\mathcal{W}$.

Some nouns contain a general meaning, i.e., "object", "entity", "thing", which will disturb the division of clusters. Intuitively, such nouns occur in most of the image-text pairs in training data and thus tend to locate near the text centers. Therefore, we compute a **uniqueness score** for each noun $\mathbf{w} \in \mathcal{W}$ as follows:

$$\rho(\mathbf{w}) = 1 - \frac{\mathbf{w}^T \mathbf{e}}{\|\mathbf{w}\|\|\mathbf{e}\|} \tag{1}$$

where $\mathbf{e} = \frac{\sum_{\mathbf{w} \in \mathcal{W}} \mathbf{w}}{|\mathcal{W}|}$ is the text center.

We set a hyperparameter γ_u to select that $\rho(\mathbf{w}) \geq \gamma_u$ as the most unique nouns by removing general worlds, resulting in a truncated noun subset $\mathcal{W}_u \in \mathcal{W}$ and $\rho(\mathbf{w}_1) \geq \rho(\mathbf{w}_2)$ holds for $\forall \mathbf{w}_1 \in \mathcal{W}_u$ and $\mathbf{w}_2 \in \mathcal{W} - \mathcal{W}_u$.

Since the nouns in $\mathcal{W}_u$ may be irrelevant to the given images $\mathcal{D}$, we further filter $\mathcal{W}_u$ according to $\mathcal{D}$. Specifically, we first perform k-means clustering on $\mathcal{U}$ to obtain c cluster centers and then select γ_r nearest nouns for each cluster center to form the final semantic set $\mathcal{T}$ and their embeddings $\mathcal{V} = h(\mathcal{T})$.

Semantic-Enhanced Pseudo-labeling

Thanks to the multi-modal pre-training models that bridge images and texts, we can connect images to semantics in an efficient way. Given the images $\mathcal{D}$ and their embeddings $\mathcal{U}$, this step aims to generate meaningful pseudo-labels according to the relationships between image embeddings $\mathcal{U}$ and semantic embeddings $\mathcal{V}$. To alleviate the above problem, we first generate c representative semantic centers $\mathcal{H}$ and then generate the pseudo-labels $\mathcal{P}$ according to both $\mathcal{U}$ and $\mathcal{H}$.

We propose three strategies for generating the c representative semantic centers $\mathcal{H}$:

1) Direct mapping. This method directly maps each image $\mathbf{u} \in \mathcal{U}$ to its nearest semantic $\mathbf{v} \in \mathcal{V}$ where the dot product is the similarity function, and forms $\mathcal{S}$ as the nearest semantic set. Then we can perform k-means clustering on $\mathcal{S}$ to obtain c cluster centers.

2) Center-based mapping. Although the first method is very simple, it is cost-intensive and may result in ambiguous nearest semantics (see Figure 2) leading to meaningless representative semantic centers $\mathcal{H}$. Intuitively, if we cluster images according to the image features, the image cluster centers are more meaningful, and mapping the image cluster centers to semantics is more appealing. Given $\mathcal{Q} = f(g(\mathcal{D}); \phi)$, we first select the top-ξ_c images for each cluster by computing a binary matrix $\mathbf{Z}$, in which $z_{il} = 1$ represents x_i is selected as the top-ξ_c samples for the l-th cluster, as follows:

$$z_{il} = \begin{cases} 1, & q_{il} \geq \kappa_l, \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

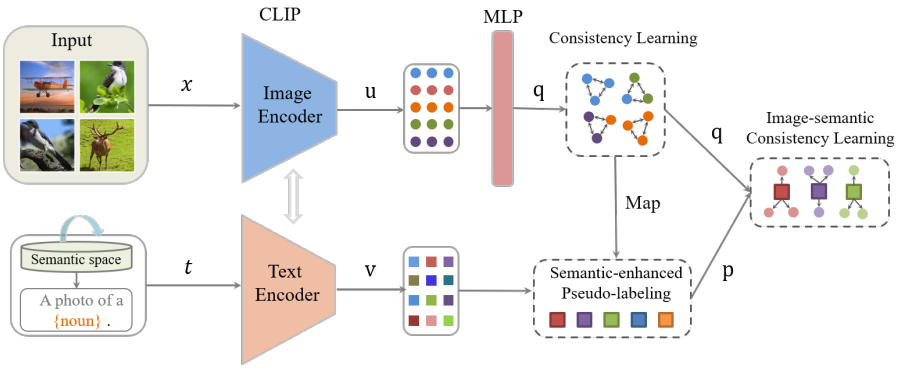


Figure 3: The framework of SIC consists of three parts: (1) Semantic space construction. (2) Semantic-enhanced pseudo-labeling. (3) Joint consistency learning. Image features and semantic features are indicated by circle and square, respectively.

where κ_l is a dynamic threshold to cut the top branch samples according to cluster assignment probabilities as the epoch evolves, which is computed as:

$$\kappa_l := \text{argtop-}\xi_c(\bar{\mathbf{q}}_l) \tag{3}$$

where $\bar{\mathbf{q}}_l$ represents the l-th column of $\mathcal{Q}$. Finally, we compute the image center $\mathcal{V}^c$ as follows:

$$\mathbf{v}_l^c = \frac{1}{\|\mathbf{z}_l\|_1} \sum z_{il} \mathbf{u}_i \tag{4}$$

After that, finding one semantic from $\mathcal{T}$ which is nearest to each image center in $\mathcal{V}^c$ results in the semantic centers $\mathcal{H}$.

3) Adjusted center-based mapping. Although the image cluster centers may map to more meaningful objects, the resulting semantic centers correspond to a set of nouns, which may limit the feasibility of the pseudo-labeling. In this method, we propose to recompute the semantic centers in $\mathcal{H}$ obtained by the second method. We first find ξ_a nearest neighborhoods for each semantic $\mathbf{h} \in \mathcal{H}$ and then recompute the centers for each semantic as the final semantic centers $\mathcal{H}$.

With the semantic centers $\mathcal{H}$, we propose an efficient way to generate the pseudo-labels. Given an image x_i, we first apply the dot product to measure the similarities between an image embedding $\mathbf{u}_i$ and c semantic centers in $\mathcal{H}$ and then conduct a softmax operation following by an argmax operation to generate pseudo-labels $\mathcal{P}$ as follows:

$$\mathbf{p}_i = \text{one-hot}\left(c, \text{argmax}_l \frac{\exp\left(\mathbf{u}_i^T \mathbf{h}_l\right)}{\Sigma_{l'}^c \exp\left(\mathbf{u}_i^T \mathbf{h}_{l'}\right)}\right) \tag{5}$$

where $\text{one-hot}(c, l)$ will generate a c-bit one-hot vector with only one 1 in the l-th position.

Joint Consistency Learning

Given an image x_i, we define its nearest neighborhood set as $\mathcal{N}_k(x_i)$, where k is a predefined parameter for the nearest neighborhoods. To learn the model $f(g(\mathcal{D}); \phi)$, we introduce the following assumptions for consistency learning:

Assumption 1 ***Local smoothness assumption****(Assumption for the consistency learning). If two images x_i and x_j are located in a local neighborhood in the low-dimensional manifold, i.e, $x_j \in \mathcal{N}_k(x_i)$, then they have similar soft cluster assignments, i.e., $\mathbf{q}_i$ and $\mathbf{q}_j$ are similar.*

Image consistency learning: according to the **local smoothness assumption**, we can learn the model $f(g(\mathcal{D}); \phi)$ by enforcing the consistency between neighborhoods in the image space with the following loss:

$$\mathcal{L}_I(f(g(\mathcal{D}); \phi)) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=rn(\mathcal{N}_k(x_i))} \log \mathbf{q}_i^T \mathbf{q}_j \tag{6}$$

where $\mathcal{N}_k(x_i)$ contains nearest neighbors of x_i and $rn(\mathcal{N}_k(x_i))$ randomly selects a sample from $\mathcal{N}_k(x_i)$ for saving the computing cost.

Image-semantic consistency learning: With the generated pseudo-labels, we perform self-supervised learning of the model $f(g(\mathcal{D}); \phi)$ with the following loss:

$$\mathcal{L}_{IS} = \frac{1}{n} \sum_{i=1}^{n} CE(\mathbf{p}_i, \mathbf{q}_i) \tag{7}$$

where $CE(.)$ is the cross entropy function.

Inspired by the k-meansNet (Peng et al. 2018), we perform k-means clustering on $\mathcal{U}$ to obtain c cluster centers as $\mathbf{R} \in \mathbb{R}^c$ to initialize the MLP parameters ϕ for reducing training time as follows:

$$\mathbf{W} = 2\tau_m \mathbf{R} \tag{8}$$

$$\mathbf{b} = \{-\tau_m \|\mathbf{h}_l\|_2^2\}_{l=1}^c \tag{9}$$

where $\mathbf{W}$ and $\mathbf{b}$ are the weight and bias of MLP, and τ_m is the temperature parameter in the MLP model.

Balance regularization: We introduce the popular negative entropy loss for the balance clustering regularization, which can prevent the model from generating empty clusters:

$$\mathcal{L}_B(f(g(\mathcal{D}); \phi)) = -\sum_{l=1}^{c} \bar{q}_l \log \bar{q}_l \tag{10}$$

where $\bar{q}_l = \frac{\sum_{i=1}^n q_{il}}{n}$ is the average cluster assignment.

The Overall Objective

The overall objective can be formulated as:

$$\min_{\phi} \mathcal{L}(f(g(\mathcal{D}); \phi)) = \min_{\phi} \mathcal{L}_I(f(g(\mathcal{D}); \phi)) + \beta \mathcal{L}_{IS}(f(g(\mathcal{D}); \phi)) + \lambda \mathcal{L}_B(f(g(\mathcal{D}); \phi)) \tag{11}$$

Algorithm 1: Semantic-Enhanced Image Clustering

Input: Images set $\mathcal{D}$, nouns set $\mathcal{W}$, neural networks $g(.)$, $h(.)$ and $f(.;\phi)$, training epoch T, cluster number c, hyperparameters γ_u and γ_r, threshold κ, nearest neighborhoods number k, trade-off parameters λ and β.
Output: Cluster assignments $\mathbf{Y}$.
Update $\mathcal{U} = g(\mathcal{D})$ and $\mathcal{V} = h(\mathcal{T})$.
Filter $\mathcal{W}$ to obtain the semantic set $\mathcal{T}$ and embeddings $\mathcal{V}$ via **Semantic Space Construction**.
Initialize ϕ^0 and $\mathcal{P}^0$.
for *t = 0 to T* **do**
  Update $\mathcal{Q}^{(t+1)} = f(g(\mathcal{D}); \phi^{(t)})$.
  Generate c representative semantic centers $\mathcal{H}$ from $\mathcal{U}$, $\mathcal{V}$ and $\mathcal{Q}^{(t+1)}$.
  Update pseudo-labels $\mathcal{P}^{t+1}$ via Eq. (5).
  Update $\phi^{(t+1)}$ by optimizing Eq. (11).
end
Output cluster assignments $\mathbf{Y}$ by $\mathbf{y}_i = \text{one-hot}\left(\text{argmax}_j q_{ij}^{(T+1)}\right)$.

where β and λ are two trade-off parameters.

Theoretical Analysis

In this part, we first analyze the convergence of our proposed method and then its expectation risk. Before analyzing, we first introduce the following assumptions

Assumption 2 ***Neighborhood Consistency Bound:*** $\forall x_i \in \mathcal{X}$, $x_j \in \mathcal{N}_k(x_i)$, $\mathbf{q}_i^T \mathbf{q}_j \in [\mu_n, 1]$.

Assumption 3 ***Prediction Confidence Bound:*** $\forall x_i \in \mathcal{X}$, $\|\mathbf{q}_i\|_\infty \leq \mu_p$.

Assumption 4 ***Neighborhood Imbalance Bound:*** *$\forall x_i \in \mathcal{X}$, x_i is in at most k' samples' (in $\mathcal{X}$) nearest neighborhoods.*

We first give the following theorem demonstrating that the optimization algorithm theoretically converges to the local optima.

Theorem 1 *Suppose that $f(.;\phi)$ and loss function $\mathcal{L}(f(g(\mathcal{D});\phi))$ are twice differential with bound gradients and Hessians. Besides, we assume that the loss function $\mathcal{L}(f(g(\mathcal{D});\phi))$ is Lipschitz smooth with constant L. Suppose that the learning rate η_ϕ satisfies $\eta_\phi = \min\{\frac{1}{L}, \frac{C}{\sqrt{T}}\}$ for some $C > 0$, such that $\frac{\sqrt{T}}{C} \geq L$. Then our proposed method can achieve $\min_{0\leq t\leq T} \mathbb{E}\left[\left\|\nabla \mathcal{L}(g(\mathcal{D});\phi^{(t)})\right\|_2^2\right] \leq \epsilon$ in $\mathcal{O}\left(1/\epsilon^2\right)$ steps, where ϵ is a very small positive real number.*

Next, we analyze the ability of our method to achieve cluster performance on unseen data. Let $\widehat{\mathcal{L}}(f(g(\mathcal{D});\phi))$ be the empirical clustering risk of our method and its expectation can be donated as $\mathcal{L}(f(g(\mathcal{X});\phi))$. The family of f is defined as $\mathcal{F}$. Recent works (Liu 2021; Li and Liu 2021; Tang and Liu 2022) establish pioneering theoretical analysis for sharper generalization bound of clustering approaches. Inspired by these studies, we obtain the following theorem by analyzing the generalization bound of our proposed method.

Theorem 2 *For any $0 < \delta < 1$, with at least probability $1-\delta$ for any $f \in \mathcal{F}$, the following inequality holds*

$$\mathcal{L}(f(g(\mathcal{X});\phi)) \leq \widehat{\mathcal{L}}(f(g(\mathcal{D});\phi)) + \frac{\tilde{c}_1}{\sqrt{n}} + \tilde{c}_2\sqrt{\frac{1}{2n}\log\frac{1}{\delta}}.$$

where $\tilde{c}_1 = 2\mu_n^{-1} + 2C\beta + 2c\lambda\log\mu_p^{-1}$ and $\tilde{c}_2 = (2 + 2k')\log\mu_n^{-1} + C\beta + 2c\lambda\log\mu_p^{-1}$. C is a constant for the function $x\log x$.

Theorem 2 shows that our proposed method, with high probability $1-\delta$, is with a bounded expected clustering risk on the unseen data. To summarize, the proposed method is theoretically guaranteed to generalize clustering tasks. Note that $\mathcal{L}(f(g(\mathcal{X});\phi))$ is inversely proportional to μ_n and μ_p which reflect the neighborhood consistency and prediction confidence, indicating that improving the neighborhood consistency and prediction confidence reduces the expected risk. Meanwhile, $\mathcal{L}(f(g(\mathcal{X});\phi))$ is proportional to k' which reflects the neighborhood overlapping, indicating that reducing the neighborhood imbalance (e.g., by setting a smaller number of neighbors k or filtering neighborhoods to reduce neighborhood imbalance) also reduces the expected risk.

Experiments and Analysis

In this section, we conduct experiments on various public benchmark datasets to evaluate our proposed method.

Experimental Setup

Datasets. We evaluated our method on five benchmark datasets, i.e. Cifar10 (Krizhevsky 2009), Cifar100-20 (Krizhevsky 2009), STL10 (Coates, Ng, and Lee 2011), ImageNet-Dogs (Chang et al. 2017b) and Tiny-ImageNet (Le and Yang 2015). A brief description of these datasets is shown in Table 1.

Dataset	Image size	#Training	#Testing	#Classes
STL10	96×96	$5,000$	$8,000$	10
Cifar10	32×32	$50,000$	$10,000$	10
Cifar100-20	32×32	$50,000$	$10,000$	20
ImageNet-Dogs	224×224	$19,500$	750	15
Tiny-ImageNet	64×64	$100,000$	$10,000$	200

Table 1: Characteristics of five benchmark datasets.

Evaluation metrics. We evaluate clustering results by three widely used metrics, including clustering Accuracy (ACC), Normalized Mutual Information (NMI) (McDaid, Greene, and Hurley 2011) and Adjusted Rand Index (ARI) (Hubert and Arabie 1985).

Dataset	η_ϕ	γ_u	γ_r	ξ_c	ξ_a	k	λ	β
STL10	1e-4	0.05	200	$0.9n/c$	20	20	5	1
Cifar10	1e-4	0.05	500	$0.9n/c$	30	20	5	0.1
Cifar100-20	1e-4	0.05	200	$0.9n/c$	20	20	5	1
ImageNet-Dogs	9e-3	0.05	1000	$0.9n/c$	50	20	5	1
Tiny-ImageNet	1e-4	0.05	200	$0.9n/c$	5	50	5	1

Table 2: The best hyper-parameters for each task. η_ϕ: learning rate, γ_u: the number of most unique nouns, γ_r: the number of nearest nouns for each image center, ξ_c: the number of the top branch samples, ξ_a: the number of nouns nearest to the image center, k: the number of nearest neighbors in image consistency learning loss. λ and β: trade-off parameters.

Implementation details. For representation learning, we used the CLIP pre-training model, whose visual and text backbones are ViT-32 (Dosovitskiy et al. 2020) and Transformer (Vaswani et al. 2017), separately. We obtained features from the image encoder of CLIP and then trained a cluster head. The cluster head is a fully connected layer with a size of $d \times c$, where d and c are the pre-training feature dimension and the number of clusters, respectively. During the training, the epoch numbers of all datasets were set to 100 with a batch size of 128. Before training, all datasets were augmented with the same method used in CLIP (Radford et al. 2021), i.e., a random square crop from resized images. The nearest neighbors were searched through Faiss Library (Johnson, Douze, and Jégou 2021). The best hyper-parameters used for five benchmark datasets are shown in Table 2.

Comparisons with State-of-the-art

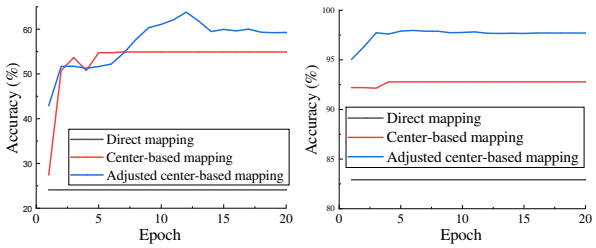
To evaluate the effectiveness of our proposed method, we compared it with 20 state-of-the-art clustering approaches on five datasets, including k-means (MacQueen 1967), SC (Zelnik-Manor 2005), NMF (Cai et al. 2009), JULE (Yang, Parikh, and Batra 2016), SAE (Ng 2011), DAE (Vincent et al. 2010), AE (Bengio et al. 2006), VAE (Kingma and Welling 2014), DEC (Xie, Girshick, and Farhadi 2016), ADC (Haeusser et al. 2018), DeepCluster (DC) (Caron et al. 2018), DAC (Chang et al. 2017a), DDC (Chang et al. 2019), DCCM (Wu et al. 2019), IIC (Ji, Vedaldi, and Henriques 2019b), PICA (Huang, Gong, and Zhu 2020), GCC (Zhong et al. 2021), CC (Li et al. 2021), SCAN(Van Gansbeke et al. 2020) and NNM (Dang et al. 2021). As shown in Table 3, different from most prior methods of training and evaluating the whole datasets on the top corner, we train and evaluate SCAN, NNM and SIC by using the train and val split respectively like SCAN (Van Gansbeke et al. 2020), which allows us to study the generalization properties of our method for novel unseen examples. The clustering results of six methods, i.e., SC (Zelnik-Manor 2005), NMF (Cai et al. 2009), AE (Bengio et al. 2006), DAE (Vincent et al. 2010) and VAE (Kingma and Welling 2014), are obtained via k-means.

Table 3 shows the clustering results of our proposed method and the state-of-the-art methods on five benchmark datasets [1]. It is clear that our proposed method outperforms all other methods on five datasets. Especially, our proposed method improves ACC, NMI and ARI by 17.2%, 25.6%, and 21.3% on the STL10 dataset, 7.7%, 10.7%, and 10.7% on the Cifar100-20 dataset and 19.2%, 10.7%, and 18.2% on the Tiny-ImageNet dataset relative to the best results of all other methods. This means that our proposed method achieves a stable superior performance.

Ablation Studies

Loss components effectiveness. We quantify the performance of loss components in our method through an ablation analysis, which consists of three losses: (a) the loss $\mathcal{L}_I$ for consistency between the image and its neighbor. (b) the loss $\mathcal{L}_{IS}$ for image-semantic consistency learning. (c) the loss $\mathcal{L}_B$ for the balance clustering regularization. Here we list the results on the Cifar10 in Table 4. Both the losses $\mathcal{L}_I$ and $\mathcal{L}_B$ play a vital role in the overall performance improvement. Combine the loss $\mathcal{L}_{IS}$ to cluster together, the performance is improved by 8.2%, 6.8% and 12.0% in terms of ACC, NMI and ARI, which indicates the effectiveness of our proposed image-semantic consistency learning.

Comparison on three semantic mapping methods. We also conduct experiments to compare the three methods for mapping images to semantic centers, i.e., **direct mapping**, **center-based mapping** and **adjusted center-based mapping**. As shown in Table 3, SIC with adjusted center-based mapping achieves the best results, and SIC with direct mapping achieves the worst results.



(a) ImageNet-Dogs. (b) STL10.

Figure 4: The accuracy of pseudo-labels as epoch evolves on ImageNet-Dogs and STL10 datasets.

We also investigate the quality of pseudo-labels generated by three methods, i.e., **direct mapping**, **center-based mapping** and **adjusted center-based mapping**. As shown in Figure 4, we can observe that SIC with adjusted center-based mapping performs best while SIC with direct mapping performs worst.

The above results verify that direct mapping each image to its nearest semantic and performing k-means to obtain centers are not good and result in low-quality pseudo-labels. Moreover, applying the adjusted centers improves the semantic centers and pseudo-labels.

[1]The clustering results (excluding those of our proposed method) are from the corresponding papers.

Dataset	STL10			Cifar10			Cifar100-20			ImageNet-Dogs			Tiny-ImageNet		
Metrics	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
k-means	19.2	12.5	6.1	22.9	8.7	4.9	13.0	8.4	2.8	10.5	5.5	2.0	2.5	6.5	0.5
SC	15.9	9.8	4.8	24.7	10.3	8.5	13.6	9.0	2.2	11.1	3.8	1.3	2.2	6.3	0.4
NMF	18.0	9.6	4.6	19.0	8.1	3.4	11.8	7.9	2.6	11.8	4.4	1.6	2.9	7.2	0.5
JULE	27.7	18.2	16.4	27.2	19.2	13.8	13.7	10.3	3.3	13.8	5.4	2.8	3.3	10.2	0.6
SAE	32.0	25.2	16.1	29.7	24.7	15.6	15.7	10.9	4.4	–	–	–	–	–	–
DAE	30.2	22.4	15.2	29.7	25.1	16.3	15.1	11.1	4.6	19.0	10.4	7.8	3.9	12.7	0.7
AE	30.3	25.0	16.1	31.4	23.4	16.9	16.5	10.0	4.7	18.5	10.4	7.3	4.1	13.1	0.7
VAE	28.2	20.0	14.6	29.1	24.5	16.7	15.2	10.8	4.0	17.9	10.7	7.9	3.6	11.3	0.6
DEC	35.9	27.6	18.6	30.1	25.7	16.1	18.5	13.6	5.0	19.5	12.2	7.9	3.7	11.5	0.7
ADC	53.0	–	–	32.5	–	–	16.0	–	–	–	–	–	–	–	–
DC	33.4	–	–	37.4	–	–	18.9	–	–	–	–	–	–	–	–
DAC	47.0	36.6	25.6	52.2	40.0	30.1	23.8	18.5	8.8	27.5	21.9	11.1	6.6	19.0	1.7
DDC	48.9	37.1	26.7	52.4	42.4	32.9	–	–	–	–	–	–	–	–	–
DCCM	48.2	37.6	26.2	62.3	49.6	40.8	32.7	28.5	17.3	38.3	32.1	18.2	10.8	22.4	3.8
IIC	59.6	49.6	39.7	61.7	51.1	41.1	25.7	22.5	11.7	–	–	–	–	–	–
PICA	71.3	61.1	53.1	69.6	59.1	51.2	33.7	31.0	17.1	35.2	35.2	20.1	9.8	27.7	4.0
GCC	78.8	68.4	63.1	85.6	76.4	72.8	47.2	47.2	30.5	52.6	49.0	36.2	13.8	34.7	7.5
CC	85.0	76.4	72.6	79.0	70.5	63.7	42.9	43.1	26.6	42.9	44.5	27.4	14.0	34.0	7.1
SCAN*	75.5(2.0)	65.4(1.2)	59.0(1.6)	81.8(0.3)	71.2(0.4)	66.5(0.4)	42.2(3.0)	44.1(1.0)	26.7(1.3)	55.6(1.5)	58.7(1.3)	42.8(1.3)	41.1(0.5)	69.4(0.3)	32.7(0.4)
SCAN†	76.7(1.9)	68.0(1.2)	61.6(1.8)	87.6(0.4)	78.7(0.5)	75.8(0.7)	45.9(2.7)	46.8(1.3)	30.1(2.1)	59.2(0.2)	60.8(0.4)	45.3(0.4)	–	–	–
SCAN†	80.9	69.8	64.6	88.3	79.7	77.2	50.7	48.6	33.3	59.3	61.2	45.7	42.0	69.8	33.2
NNM	76.8(1.2)	66.3(1.3)	59.6(1.5)	83.7(0.3)	73.7(0.5)	69.4(0.6)	45.9(0.2)	48.0(0.4)	30.2(0.4)	58.6(1.5)	60.4(0.5)	44.9(0.2)	37.8(0.1)	66.3(0.1)	27.1(0.1)
SIC^1	95.5(0.1)	92.7(0.2)	91.1(0.2)	78.3(0.1)	74.3(0.1)	66.9(0.1)	51.3(0.1)	53.9(0.1)	36.8(0.1)	59.0(0.2)	57.7(1.8)	41.1(3.2)	55.7(0.8)	77.4(0.1)	44.9(0.6)
SIC^2	96.7(0.1)	93.7(0.1)	93.2(0.1)	91.8(0.1)	83.4(0.1)	83.1(0.1)	54.0(0.1)	54.4(0.4)	38.6(0.4)	61.8(1.1)	63.9(1.9)	49.8(1.4)	61.0(0.2)	80.4(0.1)	51.2(0.2)
SIC^3	98.1(0.1)	95.3(0.1)	95.9(0.1)	92.6(0.1)	84.7(0.1)	84.4(0.1)	58.3(0.1)	59.3(0.1)	43.9(0.1)	69.7(1.1)	69.0(1.6)	55.8(1.5)	60.2(0.3)	79.4(0.1)	49.4(0.2)
SIC	**98.1**	**95.4**	**95.9**	**92.7**	**84.8**	**84.6**	**58.4**	**59.3**	**44.0**	**71.3**	**71.8**	**58.6**	**61.2**	**80.5**	**51.4**

Table 3: State-of-the-art comparison results on five benchmarks, including the averaged results of 5 different runs with standard deviation and the best model. The methods evaluation is divided into the whole dataset (top corner) and split datasets (bottom corner). We evaluated our proposed method on split datasets. SIC$^{1-3}$ represent SIC with direct mapping, center-based mapping and adjusted center-based mapping, respectively. The best results are shown in boldface.

Setup	ACC	NMI	ARI
w/o $\mathcal{L}_{IS}$	84.4 ± 0.5	77.9 ± 0.3	72.4 ± 0.5
w/o$\mathcal{L}_I$	71.2 ± 0.1	68.2 ± 0.2	59.2 ± 0.3
w/o$\mathcal{L}_B$	70.3 ± 7.2	74.6 ± 2.5	58.8 ± 6.0
SIC	$\mathbf{92.6 \pm 0.1}$	$\mathbf{84.7 \pm 0.1}$	$\mathbf{84.4 \pm 0.1}$

Table 4: Ablation studies of our method on Cifar10 dataset.

Methods (ACC)	STL10	Cifar10	ImageNet-Dogs
CLIP+zero-shot	95.7	80.0	34.1
CLIP+k-means	94.6±0.1	75.3±0.1	39.8±3.9
SIC	**98.1±0.1**	**92.6±0.1**	**69.7±1.1**

Table 5: Ablation studies of our method compared to "CLIP+zero-shot" and "CLIP+k-means".

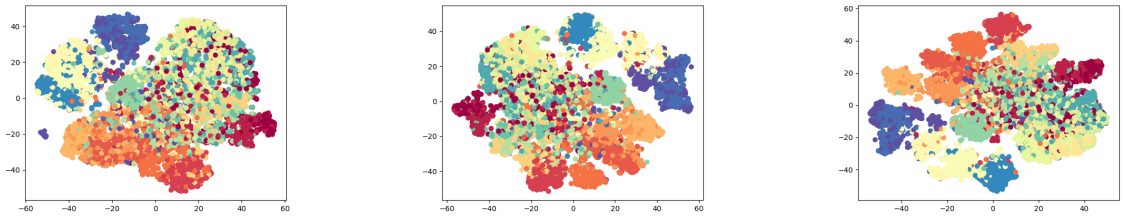
Compared to relative results of CLIP. To display the clustering power of our model, we compare SIC with "CLIP+zero-shot" and "CLIP+k-means" on the STL10, Cifar10, and ImageNet-Dogs datasets. "CLIP+zero-shot" uses the given class names in each dataset to directly classify images with CLIP, and "CLIP+k-means" performs k-means clustering on image embeddings obtained by the image encoder in CLIP. In Table 5, it is clear that SIC outperforms the other two methods, indicating that our method can better utilize CLIP to uncover image clusters without class names.

Visualization of learned image features. Figure 5 visualizes the image features obtained by CLIP, image consistency learning (before softmax), and SIC (before softmax) by t-SNE on the Cifar100-20 dataset. We can observe ambiguous cluster structures from the image features obtained by CLIP. Although image consistency learning improves image embeddings, we also observe ambiguous cluster structures. However, with our proposed method, we can observe the clearest structures.

Sensitivity Analysis

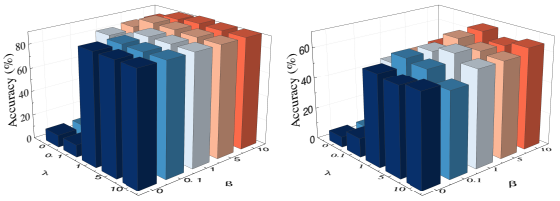
Sensitivity on trade-off parameters λ and β. We study the influence of trade-off parameters λ and β, where β helps to separate the visually similar but semantically different images and λ helps prevent the model into a trivial solution. We set $\lambda, \beta \in [0, 0.1, 1, 5, 10]$ to show the sensitivity results in Figure 6. In general, decreasing λ causes performance degradation, and increasing β improves performance.

Sensitivity on hyperparameters γ_u and γ_r. In our method, γ_u and γ_r are used to select proper nouns from WordNet by removing the general words. As shown in Figure 7, we can observe that decreasing γ_u and increasing γ_r improves the performance first, then does not improve the performance too much, indicating that removing general worlds can re-



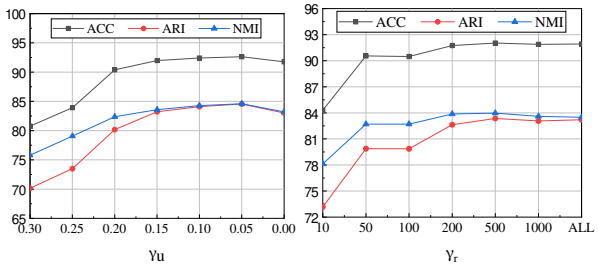
(a) CLIP (b) Image consistency learning (c) SIC

Figure 5: t-SNE visualization of learned image features from CLIP, image consistency learning, and SIC on the Cifar100-20.



(a) (λ,β) on Cifar10. (b) (λ,β) on ImageNet-Dogs.

Figure 6: Sensitivity analysis of λ and β.



(a) γ_u. (b) γ_r.

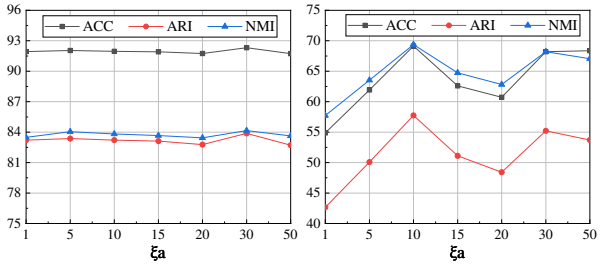
Figure 7: Sensitivity analysis of γ_u and γ_r on Cifar10.

duce the computing cost without performance degeneration too much. We also observe that decreasing μ_u from 0.05 to 0 causes performance degradation, indicating that removing general words is necessary.

Sensitivity on hyperparameters ξ_a and ξ_c. ξ_a is used to adjust the semantic centers and ξ_c is used to select the top branch samples. Figures 8 and 9 show the sensitivity results on both ξ_a and ξ_c, respectively. We can observe different sensitives of ξ_a and ξ_c on different datasets. For example, ξ_a and ξ_c do not affect the performance too much on the Cifar10 dataset, but affect too much on the ImageNet-Dogs dataset.

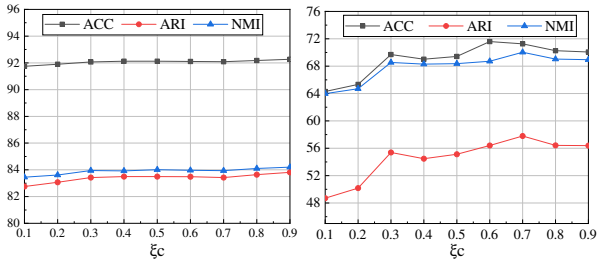
Conclusion

This paper proposes a novel image clustering **SIC** which utilizes the visual-language pre-training model CLIP to compensate the semantic information for better image clustering. We propose efficient methods to map images to a proper



(a) ξ_a on Cifar10. (b) ξ_a on ImageNet-Dogs.

Figure 8: Sensitivity analysis of ξ_a.



(a) ξ_c on Cifar10. (b) ξ_c on Imagenet-dogs.

Figure 9: Sensitivity analysis of ξ_c.

semantic space and cluster images from both image and semantic spaces. Theoretical results show that SIC can converge and reveal that the expected risk of SIC is affected by the models' performance in terms of neighborhood consistency and prediction confidence. The imbalance of the constructed neighborhoods also affects the expected risk of SIC. Experimental results show that our method outperforms 20 state-of-the-art and zero-shot learning with CLIP, enabling its wide potential applications. However, the pseudo-labels generated in our method may be suboptimal, so we will study new methods to generate better pseudo-labels. It is deserved to extend our theoretical results to self-supervised learning.

Acknowledgments

This work is jointly supported by Major Project of the New Generation of Artificial Intelligence (No. 2018AAA0102900); in part by NSFC under Grant no. 92270122 and no. 62206179; and in part by the Shenzhen Research Foundation for Basic Research, China, under Grant JCYJ20210324093000002.

References

Bengio, Y.; Lamblin, P.; Popovici, D.; and Larochelle, H. 2006. Greedy layer-wise training of deep networks. In *Proceedings of NIPS 2006*, 153–160.

Cai, D.; He, X.; Wang, X.; Bao, H.; and Han, J. 2009. Locality Preserving Nonnegative Matrix Factorization. In *Proceedings of IJCAI 2009*, 1010–1015.

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

Chang, J.; Guo, Y.; Wang, L.; Meng, G.; Xiang, S.; and Pan, C. 2019. Deep Discriminative Clustering Analysis. arXiv:1905.01681.

Chang, J.; Wang, L.; Meng, G.; Xiang, S.; and Pan, C. 2017a. Deep Adaptive Image Clustering. In *Proceedings of ICCV 2017*, 5880–5888.

Chang, J.; Wang, L.; Meng, G.; Xiang, S.; and Pan, C. 2017b. Deep Adaptive Image Clustering. In *IEEE International Conference on Computer Vision, ICCV 2017*, 5880–5888.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120. Springer.

Coates, A.; Ng, A.; and Lee, H. 2011. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. In *Proceedings of AISTATS 2011*, volume 15, 215–223.

Dang, Z.; Deng, C.; Yang, X.; Wei, K.; and Huang, H. 2021. Nearest Neighbor Matching for Deep Clustering. In *Proceedings of CVPR 2021*, 13693–13702.

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

Gupta, D.; Ramjee, R.; Kwatra, N.; and Sivathanu, M. 2020. Unsupervised Clustering using Pseudo-semi-supervised Learning. In *Proceedings of ICLR 2020*.

Haeusser, P.; Plapp, J.; Golkov, V.; Aljalbout, E.; and Cremers, D. 2018. Associative deep clustering: Training a classification network with no labels. In *Proceedings of GCPR 2018*, 18–32.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, J.; Gong, S.; and Zhu, X. 2020. Deep Semantic Clustering by Partition Confidence Maximisation. In *Proceedings of CVPR 2020*, 8846–8855.

Hubert, L.; and Arabie, P. 1985. Comparing partitions. *Journal of Classification*, 2(1): 193–218.

Ji, P.; Zhang, T.; Li, H.; Salzmann, M.; and Reid, I. 2017. Deep Subspace Clustering Networks Pan. In *Proceedings of NeurIIPS 2017*, 23–32.

Ji, X.; Vedaldi, A.; and Henriques, J. 2019a. Invariant Information Clustering for Unsupervised Image Classification and Segmentation. In *Proceedings of ICCV 2019*, 9864–9873.

Ji, X.; Vedaldi, A.; and Henriques, J. F. 2019b. Invariant Information Clustering for Unsupervised Image Classification and Segmentation. In *Proceedings of ICCV 2019*, 9864–9873.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 4904–4916. PMLR.

Jin, C.; Mao, W.; Zhang, R.; Zhang, Y.; and Xue, X. 2015. Cross-modal image clustering via canonical correlation analysis. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 151–159.

Johnson, J.; Douze, M.; and Jégou, H. 2021. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7(3): 535–547.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. arXiv:1312.6114.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. *Master's thesis, University of Tront*.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

Li, G.; Duan, N.; Fang, Y.; Gong, M.; and Jiang, D. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11336–11344.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Li, S.; and Liu, Y. 2021. Sharper generalization bounds for clustering. In *International Conference on Machine Learning*, 6392–6402. PMLR.

Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; and Wang, H. 2020b. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*.

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

Liu, Y. 2021. Refined Learning Bounds for Kernel and Approximate k-Means. *Advances in Neural Information Processing Systems*, 34: 6142–6154.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observation. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistica and Probability*, 281–297.

Mahon, L.; and Lukasiewicz, T. 2021. Selective Pseudo-Label Clustering. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)*, 158–178. Springer.

McDaid, A. F.; Greene, D.; and Hurley, N. 2011. Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv:1110.2515*.

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

Ng, A. 2011. Sparse autoencoder. *CS294A Lecture notes*.

Park, S.; Han, S.; Kim, S.; Kim, D.; Park, S.; Hong, S.; and Cha, M. 2021. Improving Unsupervised Image Clustering With Robust Learning. In *Proceedings of CVPR 2021*, 12278–12287.

Peng, X.; Tsang, I. W.; Zhou, J. T.; and Zhu, H. 2018. k-meansnet: When k-means meets differentiable programming. *arXiv preprint arXiv:1808.07292*.

Qi, D.; Su, L.; Song, J.; Cui, E.; Bharti, T.; and Sacheti, A. 2020. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*.

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

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Shaham, U.; and Stanton, K. 2018. SpectralNet: Spectral Clustering using Deep Neural Networks. In *ICLR 2018*, 1–20.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tang, H.; and Liu, Y. 2022. Deep Safe Incomplete Multi-view Clustering: Theorem and Algorithm. In *International Conference on Machine Learning*, 21090–21110. PMLR.

Tian, K.; Zhou, S.; and Guan, J. 2017. DeepCluster: A General Clustering Framework Based on Deep Learning. In *Proceedings of ECML PKDD 2017*, 809–825.

Van Gansbeke, W.; Vandenhende, S.; Georgoulis, S.; Proesmans, M.; and Van Gool, L. 2020. SCAN: Learning to Classify Images Without Labels. In *Proceedings of ECCV 2020*, 268–285.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

Wang, Q.; Lian, H.; Sun, G.; Gao, Q.; and Jiao, L. 2020. ICMSC: Incomplete cross-modal subspace clustering. *IEEE Transactions on Image Processing*, 30: 305–317.

Wu, J.; Long, K.; Wang, F.; Qian, C.; Li, C.; Lin, Z.; and Zha, H. 2019. Deep Comprehensive Correlation Mining for Image Clustering. In *Proceedings of ICCV 2019*, 8149–8158.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised Deep Embedding for Clustering Analysis. In *Proceedings of ICML 2016*, 478–487.

Yang, B.; Fu, X.; Sidiropoulos, N. D.; and Hong, M. 2017. Towards K-means-friendly Spaces: Simultaneous Deep Learning and Clustering. In *Proceedings of ICML 2017*, volume 70, 3861–3870.

Yang, J.; Parikh, D.; and Batra, D. 2016. Joint Unsupervised Learning of Deep Representations and Image Clusters. In *Proceedings of CVPR 2016*, 5147–5156.

Yang, S. T.; Huang, K.-H.; and Howe, B. 2021. JECL: Joint Embedding and Cluster Learning for Image-Text Pairs. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 8344–8351. IEEE.

Zelnik-Manor, L. 2005. Self-tuning spectral clustering. In *Proceedings of NIPS 2005*, volume 17, 1601–1608.

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