Multi-Level Cross-Modal Alignment for Image Clustering

Liping Qiu[*], **Qin Zhang**[*], **Xiaojun Chen**[†], **Shaotian Cai**

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

Abstract

Recently, the cross-modal pretraining model has been employed to produce meaningful pseudo-labels to supervise the training of an image clustering model. However, numerous erroneous alignments in a cross-modal pretraining model could produce poor-quality pseudo labels and degrade clustering performance. To solve the aforementioned issue, we propose a novel **Multi-Level Cross-Modal Alignment** method to improve the alignments in a cross-modal pretraining model for downstream tasks, by building a smaller but better semantic space and aligning the images and texts in three levels, i.e., instance-level, prototype-level, and semantic-level. Theoretical results show that our proposed method converges, and suggests effective means to reduce the expected clustering risk of our method. Experimental results on five benchmark datasets clearly show the superiority of our new method.

Introduction

Image clustering which groups images into different clusters without labels is an essential task in unsupervised learning. Many methods are proposed to utilize the large-scale pre-training models such as Resnet (He et al. 2016) or ViT (Dosovitskiy et al. 2020) to extract high-quality representations for image clustering (Ji, Vedaldi, and Henriques 2019; Li et al. 2021; Zhong et al. 2021; Wu et al. 2019; Van Gansbeke et al. 2020; Dang et al. 2021). Then, to unsupervised classification models, multiple indirect loss functions are used (*e.g.* sample relations (Chang et al. 2017a), invariant information (Ji, Vedaldi, and Henriques 2019; Li et al. 2021), mutual information (Wu et al. 2019) and entropy (Huang, Gong, and Zhu 2020; Van Gansbeke et al. 2020; Li et al. 2021). However, as pointed out in (Cai et al. 2023), the aforementioned techniques have difficulties in handling examples that are semantically different but visually comparable by focusing only on images.

Recently, many vision-language pre-training (VLP) models have been developed to align images and texts into a unified semantic space (Li et al. 2019; Chen et al. 2020; Ramesh et al. 2021; Li et al. 2020; Radford et al. 2021; Jia et al.
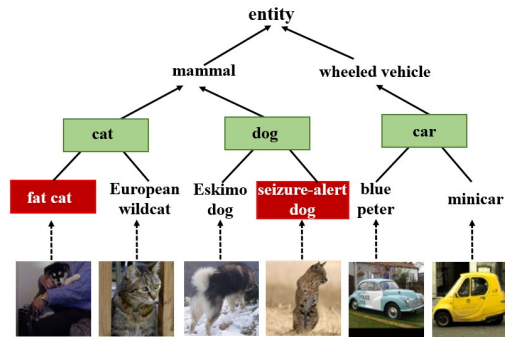
---

[*]These authors contributed equally.

[†]Corresponding author.

Figure 1: The nearest noun for images in STL10 from Word-Net, where the image and text embeddings are obtained via CLIP. The green words correspond to the correct alignments, while the red words indicate incorrect alignments.

2021). To utilize the VLP models for image clustering, Cai et al. (Cai et al. 2023) proposed to use CLIP (Radford et al. 2021) to produce meaningful pseudo-labels and achieved significant improvements on a wide range of datasets in comparison with conventional image clustering methods. Li et al. (Li, Savarese, and Hoi 2022) also used CLIP for a zero-shot image classification task. The success of these methods suggests a promising direction for image clustering. However, as depicted in Figure 1, there are instances where the alignments between images and texts in CLIP may be incorrect for downstream tasks, resulting in subpar pseudo labels and poor clustering performance. SIC (Cai et al. 2023) simply uses CLIP to obtain the embeddings of images and texts and cannot deal with incorrect alignments. While MUST (Li, Savarese, and Hoi 2022) strives to optimize the image encoder in CLIP, its efficiency is hampered by using the pretraining task to update the image encoder, resulting in a slow process.

To address the above problem, we propose a novel method, namely **Multi-Level Cross-Modal Alignment (MCA)**, an efficient way to improve the alignments between images and texts in CLIP for clustering tasks. In general, our main contributions are as follows:

- We propose to use the hierarchical structure in WordNet (see Figure 1) to filter irrelevant words and construct a

smaller but better semantic space, thus reducing the affection of unrelated nouns for clustering. Our experimental findings demonstrate that it can reduce the number of words by up to 60% and significantly enhance clustering performance compared to SIC.

- We propose to optimize both image and text embeddings for downstream tasks, by aligning the images and texts at three levels, i.e., instance-level, prototype-level, and semantic-level. Our proposed method can better fix the incorrect alignments in CLIP for downstream tasks when compared to SIC and MUST.

- Theoretical findings demonstrate that our proposed method converges at a sublinear rate and offers effective strategies for lowering the expected clustering risk of our method. These findings will provide valuable guidance for the design of new image clustering methods.

- Experimental results on five benchmark datasets clearly show the superiority of our new method, especially when dealing with complex clusters.

Related Work

The early deep clustering methods simply combine representation learning and shallow clustering (Xie, Girshick, and Farhadi 2016; Yang et al. 2017; Tian, Zhou, and Guan 2017; Shaham and Stanton 2018). With the rapid development of the pre-training paradigm, many methods employ large-scale pre-training models such as Resnet (He et al. 2016) or ViT (Dosovitskiy et al. 2020) to extract high-quality representations and train a classification model, by maximizing the consistency between each image and its augmentations/neighbors (Ji, Vedaldi, and Henriques 2019; Li et al. 2021; Zhong et al. 2021; Wu et al. 2019; Van Gansbeke et al. 2020; Zhong et al. 2021; Dang et al. 2021), or generating pseudo-labels (Wu et al. 2019; Van Gansbeke et al. 2020). However, as pointed out in (Cai et al. 2023), it is challenging for the aforementioned techniques to handle examples that are semantically different but visually comparable by only accessing visual information in images.

Cross-modal clustering has made significant progress in recent years, which usually learns a shared subspace such that the mutual agreement between multiple modalities is maximized, by Canonical Correlation Analysis (CCA) (Gao et al. 2020) or mutual information optimization (Mao et al. 2021). However, these methods require image-text pairs as input, which may be cost-intensive to collect in real applications.

Recently, vision-language pre-training (VLP) models that align multi-modal data in common feature space by different pre-training tasks have been proposed. For example, VisualBert (Li et al. 2019), UNITER (Chen et al. 2020) and DALL-E (Ramesh et al. 2021) use language-based training strategies, including mask LM (Language Modeling) such as Masked Language/Region Modeling, or autoregressive LM such as image caption and text-grounded image generation. CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021) utilize cross-modal contrastive learning to align the visual and textual information into a unified semantic space.

Since VLP captures the relationships among images and texts (low-level semantics), it is natural to utilize VLP models to compensate for the semantic information for better image clustering. Cai et al. (Cai et al. 2023) proposed to use CLIP (Radford et al. 2021) to generate meaningful pseudo-labels for image clustering. Li et al. (Li, Savarese, and Hoi 2022) also proposed to use CLIP for zero-shot image classification tasks.

Notation and Problem Definition

Suppose we have an image dataset $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ with n instances sampled i.i.d. from input space $\mathcal{D}$, we can obtain the embeddings of these images as $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ where $\mathbf{u}_i = e_I(x_i) \in \mathbb{R}^{d \times 1}$ is obtained via the image encoder $e_I(.)$ of CLIP, where d is the embedding dimension. To capture the semantic meaning of these images, we first introduce a noun vocabulary $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}$ that includes m noun phrases sampled from WordNet (Miller 1995). Then we can obtain the embeddings of these m words as $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\}$ where $\mathbf{v}_i = e_T(s_i) \in \mathbb{R}^{d \times 1}$, s_i is a sentence like "A photo of a $\{t_i\}$" and $e_T(.)$ is the text encoder of CLIP. Let c be the number of categories, our goal is to group the images in $\mathcal{X}$ into c clusters with the help of CLIP. Let $f_I(e_I(\mathcal{X}); \phi)$ denotes the image classification network with parameters ϕ that maps an image x_i into a soft cluster assignment probability vector $\mathbf{q}_i \in \mathbb{R}^{c \times 1}$, and $f_S(e_T(\mathcal{T}); \theta)$ denotes the text classification network with parameters θ that maps a word t_i into a soft cluster assignment probability vector $\mathbf{p}_i \in \mathbb{R}^{c \times 1}$. Notably, e_I and e_T in CLIP are kept frozen during the training process.

Method

In this paper, we propose our method shown in Figure 2. This new method mainly consists of three components: 1) **Semantic space construction** builds a proper semantic space $\mathcal{T}$, 2) **Image consistency learning** performs the consistency learning in image space, and 3) **Multi-level cross-modal alignment** aligns images and texts at three distinct levels: instance-level, prototype-level, and semantic-level.

Semantic Space Construction

Constructing a proper semantic space $\mathcal{T}$ from WordNet (Miller 1995) such that the images can be well represented by the words in $\mathcal{T}$ is very important for image clustering, because too small $\mathcal{T}$ may lose important relevant words but too large $\mathcal{T}$ may contain too many irrelevant noisy words. In this step, we first build a candidate semantic space $\mathcal{W}$ as 82,000 nouns in the WordNet dataset (Miller 1995). Considering an image dataset usually makes up a small part of semantics, we propose a two-step filtering strategy to construct a proper semantic space for an image dataset: 1) **Uniqueness-based filtering** selects γ_r nearest words for each of c image cluster centers obtained by $k$means of the most unique nouns whose uniqueness scores (Cai et al. 2023) are greater than a given hyperparameter ρ_u. 2) **Hierarchy-based filtering** employs the hierarchical structure in WordNet to further filter $\mathcal{W}_c$ to form
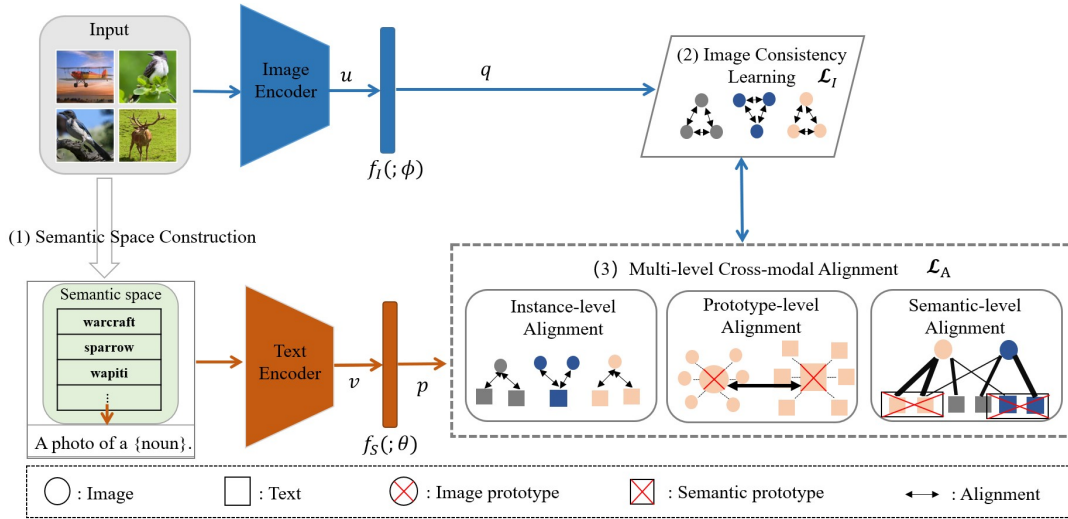
Figure 2: The framework of MCA consists of three parts: (1) Semantic space construction. (2) Image consistency learning (3) Multi-level cross-modal alignment. The thickness of lines in adaptive instance-level alignment reflects the magnitude of attention scores.

the final semantic space $\mathcal{T}$. Let $\mathcal{T} = \emptyset$. Given an image x_i, we find its nearest noun $w_i \in \mathcal{W}_c$ and search its hierarchical structure from WordNet (Miller 1995) to form a hierarchical semantic tree. In general, the words in the lower layers provide more fine-grained information for distinguishing the images, while the words in the higher layers may be useless for clustering. Figure 1 shows, for example, "mammal" is the common parent of "dog" and "cat" and cannot distinguish the images in "dog" or "cat". Therefore, we propose a hierarchy-based filtering strategy that filters out the top γ_h levels (excluding root node) and adds each of the remaining words into $\mathcal{T}$ if it is also in $\mathcal{W}_c$.

Image Consistency Learning

Intuitively, an image and its nearest images may have similar soft cluster assignments. Therefore, we propose the following loss function for image consistency learning:

$$
\begin{aligned}
\mathcal{L}_I(f_I(e_I(\mathcal{X}); \phi)) = & -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=rn(\mathcal{N}_{k_I}^I(x_i))} \log \mathbf{q}_i^T \mathbf{q}_j \\
& - \eta \sum_{l=1}^{c} \bar{q}_l \log \bar{q}_l
\end{aligned}
\tag{1}
$$

where $\mathcal{N}_{k_I}^I(x_i)$ contains k_I nearest images of x_i and $rn(\mathcal{N}_{k_I}^I(x_i))$ randomly selects a sample from $\mathcal{N}_{k_I}^I(x_i)$. The second item is the popular negative entropy loss for preventing trivial solutions that most samples belong to a small proportion of clusters, where $\bar{q}_l = \frac{\sum_{i=1}^{n} q_{il}}{n}$ is the average cluster assignment. η is a trade-off parameter.

Multi-Level Cross-Modal Alignment

When using a cross-modal pretraining model for image clustering, the main challenge is to rectify incorrect alignments

between images and words in image data. In this paper, we propose a novel **Multi-Level Cross-Modal Alignment** method for this task, which is shown in Figure 2. Specifically, our method employs a three-level alignment approach. Firstly, at the **Instance-level Alignment**, each image is aligned with its neighboring texts. Secondly, the **Prototype-level Alignment** aligns each image prototype with its nearest text prototype. Lastly, at the **Semantic-level Alignment**, each image is aligned with its neighboring texts in the semantic space. The detailed descriptions of these three alignment processes are provided below.

Instance-level alignment: Given an image x_i and its neighboring texts $\mathcal{N}_{k_S}^S(x_i)$, we propose the following contrastive loss function to facilitate the alignment:

$$
\mathcal{L}_{ia} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=rn(\mathcal{N}_{k_S}^S(x_i))} \log \frac{exp(\mathbf{q}_i^T \mathbf{p}_j / \tau_{ia})}{\sum_{l=1, l \neq j}^{m} exp(\mathbf{q}_i^T \mathbf{p}_l / \tau_{ia})}
\tag{2}
$$

where τ_{ia} is a tempreature parameter.

Prototype-level alignment: Instancel-level alignment may be affected by noisy neighborhood relationships, so we further propose to align images and texts at prototype level which is more robust to noisy texts. We first compute an image prototype set $\mathcal{H}^I$, where $\mathbf{h}_l^I \in \mathcal{H}^I$ is computed as $\mathbf{h}_l^I = \frac{1}{\|\mathbf{q}_l\|_1} \sum_{i=1}^{n} q_{il} \mathbf{u}_i$. Then, for each image prototype $\mathbf{h}_l^I \in \mathcal{H}^I$, we can identify the word in $\mathcal{T}$ that is closest to $\mathbf{h}_l^I$ and finally construct a prototype set $\mathcal{H}^S$. To further improve the prototypes in $\mathcal{H}^S$, we find k_p nearest neighborhoods for each $\mathbf{h}_l^S \in \mathcal{H}^S$ to compute the prototype of these neighborhoods and replace $\mathbf{h}^S$ to update $\mathcal{H}^S$.

Finally, to align images and texts at the prototype level,

we propose the following loss function:

$$\mathcal{L}_{pa} = -\frac{1}{c} \sum_{j=1}^{c} \log \frac{exp(f_I(\mathbf{h}_j^I, \phi)^T f_S(\mathbf{h}_j^S, \theta)/\tau_{pa})}{\sum_{l=1, l \neq j}^{c} exp(f_I(\mathbf{h}_j^I, \phi)^T f_S(\mathbf{h}_l^S, \theta)/\tau_{pa})}$$

(3)

where τ_{pa} is a tempreature parameter.

Semantic-level alignment: Given an image x_i and its neighbouring texts $\mathcal{N}_{k_S}^{S}(x_i)$, we let $\mathbf{v}_j = e_T(s_j)$ be the the embeddings of $t_j \in \mathcal{N}_{k_S}^{S}(x_i)$ and s_j is a sentence like "A photo of a $\{t_j\}$", and $\mathbf{p}_j = f_S(\mathbf{v}_j, \theta)$. It makes sense that the neighboring texts of an image can help determine the cluster assignment of this image. Note that the alignment relationships between images and texts may vary in different downstream tasks, we propose to use the attention mechanism (Vaswani et al. 2017) to quantify the correlations between an image and its neighboring texts. Specifically, we compute $\mathbf{p}_i'$ as the weighted combination of the neighboring texts' assignments as:

$$\begin{aligned} \mathbf{p}_i' &= f_A(\mathbf{u}_i, \mathbf{V}^S, \mathbf{P}^S; \mathbf{W}^I, \mathbf{W}^S) \\ &= \sum_{j \in \mathcal{N}_{k_S}^{S}(x_i)} softmax((\mathbf{W}^I \mathbf{u}_i)^T \mathbf{W}^S \mathbf{v}_j) \mathbf{p}_j \end{aligned}$$

(4)

where $f_A(\cdots; \mathbf{W}^I, \mathbf{W}^S)$ is the attention network to quantify the correlations between an image and its neighboring texts and $\mathbf{W}^I, \mathbf{W}^S \in \mathbb{R}^{d \times d}$ are two parameter matrices. Then we use the argmax operation to generate one-hot pseudo-label for x_i as:

$$\mathbf{q}_i' = \text{one-hot}\,(c, \text{argmax}_l\, p_{il}')$$

(5)

where one-hot(c, l) generates a c-bit one-hot vector with the l-th element as 1.

Here, $\mathbf{q}_i'$ can be considered as the semantic cluster assignment of x_i. Therefore, we perform alignment for each image x_i by aligning the semantic cluster assignment $\mathbf{q}_i'$ to the image cluster assignment $\mathbf{q}_i$ with the following loss function:

$$\mathcal{L}_{sa} = \frac{1}{n} \sum_{i=1}^{n} CE(\mathbf{q}_i, \mathbf{q}_i')$$

(6)

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

The overall alignment loss function is:

$$\begin{aligned} &\mathcal{L}_A(f_I(e_I(\mathcal{X}); \phi), f_S(e_T(\mathcal{T}); \theta), f_A(\cdots; \mathbf{W}^I, \mathbf{W}^S)) \\ &= \mathcal{L}_{ia} + \lambda_{pa} \mathcal{L}_{pa} + \lambda_{sa} \mathcal{L}_{sa} \end{aligned}$$

(7)

where λ_{pa} and λ_{sa} are two trade-off parameters.

The Overall Objective

$\mathcal{L}_A(f_I(e_I(\mathcal{X}); \phi), f_S(e_T(\mathcal{T}); \theta), f_A(\cdots; \mathbf{W}^I, \mathbf{W}^S))$ denote as $\mathcal{L}(g(\mathcal{S}; \varphi))$ for simplicity, where $\mathcal{S} = (\mathcal{X}, \mathcal{T})$, $\varphi = (\phi, \theta, \mathbf{W}^I, \mathbf{W}^S)$ and g consists of f_I, f_S and f_A. Finally, the overall objective can be formulated as

$$\mathcal{L}(g(\mathcal{S}; \varphi)) = \mathcal{L}_I(g(\mathcal{S}; \varphi)) + \lambda_a \mathcal{L}_A(g(\mathcal{S}; \varphi))$$

(8)

where λ_a is a trade-off parameter.

Theoretical Analysis

In this part, we first analyze the convergence of our proposed method and then analyze its expected clustering risk. We first introduce the following assumptions: **Image Neighborhood Consistency Bound:** $\forall x_i \in \mathcal{X}, x_j \in \mathcal{N}_{k_I}^{I}(x_i)$, $\mathbf{q}_i^T \mathbf{q}_j \in [\mu_I, 1]$. **Cross-modal Neighborhood Consistency Bound:** $\forall x_i \in \mathcal{X}, t_j \in \mathcal{N}_{k_S}^{S}(x_i), \mathbf{q}_i^T \mathbf{p}_j \in [\mu_C, 1]$. **Image Prediction Confidence Bound:** $\forall x_i \in \mathcal{X}, \|\mathbf{q}_i\|_\infty \leq \mu_p$. **Image Neighborhood Imbalance Bound:** $\forall x_i \in \mathcal{X}, x_i$ is in at most k_I' samples' (in $\mathcal{X}$) nearest neighborhoods.

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

Suppose that $g(\mathcal{S}; \varphi)$ is twice differential with bounded gradients and Hessians, and $\mathcal{L}(g(\mathcal{S}; \varphi))$ has L-Lipschitz continuous gradient. Suppose that the learning rate η_φ satisfies $\eta_\varphi = \min\{1, \frac{C}{\sqrt{T}}\}$ for some $C > 0$, such that $\frac{C}{\sqrt{T}} \geq L$. Then our proposed method can achieve $\min_{0 \leq t \leq T} \mathbb{E}\left[\left\|\nabla \mathcal{L}(g(\mathcal{S}); \varphi^{(t)})\right\|_2^2\right] \leq \epsilon$ 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 $\hat{\mathcal{L}}_n(g)$ be the empirical clustering risk of MCA and its expectation can be denoted as $\mathcal{L}(g)$. The family of g is defined as $\mathcal{G}$. Then we can obtain the following theorem by analyzing the generalization bound of our proposed method.

Suppose $f_I(.; \phi)$ is Lipschitz smooth with constant L_I and $\|\mathbf{u}\|_\infty \leq M_u$. Suppose $\beta(\frac{1}{n} \sum_{i=1}^{n} q_{il} \mathbf{u}_i) = f_I(\mathbf{h}^I, \phi)^T f_S(\mathbf{h}^S, \theta)$ is L_{IS}-Lipschitz continious, where $\mathbf{h}_l^I$ and $\mathbf{h}^S$ are computed according to the method in Section . For any $0 < \delta < 1$, we can guarantee that with a probability of at least $1 - \delta$ for any $g \in \mathcal{G}$, the following inequality holds.

$$\mathcal{L}(g) \leq \widehat{\mathcal{L}}_n(g) + \frac{\tilde{c}_1}{\sqrt{n}} + \tilde{c}_2 \sqrt{\frac{1}{2n} \log \delta^{-1}} + \frac{2dL_{IS}M_u}{n\tau_{pa}}.$$

where $\tilde{c}_1 = 2\mu_I^{-1} + 2\eta C + 2\lambda_a m/\tau_{ia} + 2\lambda_a \lambda_{pa} dL_{IS} M_u/\tau_{pa} + 2\lambda_a \lambda_{sa} c \log \mu_p^{-1}$ *and* $\tilde{c}_2 = (2 + 2k_I') \log \mu_I^{-1} + \eta C + \frac{2\lambda_a(1-\mu_C)}{\tau_{ia}} + \lambda_a \lambda_{pa} \frac{dcL_I M_u^2}{\tau_{pa}} + 2\lambda_a \lambda_{sa} c \log \mu_p^{-1}$ *are constants dependent on* $\{n, m, \mu_I, \mu_C, \mu_p, k_I', c, L_{IS}, L_I, M_u, d, C\}$. C *is a constant.*

Theorem shows that our proposed method, with high probability $1 - \delta$, is with a bounded expected clustering risk on the unseen data. To summarize, the expected clustering risk of MCA is theoretically guaranteed in clustering tasks. Note that the margin $\mathcal{L}(g) - \widehat{\mathcal{L}}_n(g)$ is inversely proportional to μ_I and μ_C which reflect the neighborhood consistency in both image domain and cross-domain, and μ_p which reflects the prediction confidence, indicating that improving the neighborhood consistency in both image domain and cross-domain and prediction confidence reduces the expected risk of MCA. Meanwhile, the margin $\mathcal{L}(g) - \widehat{\mathcal{L}}_n(g)$

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
TCL	86.8	79.9	75.7	88.7	81.9	78.0	53.1	52.9	35.7	64.4	62.3	51.6	–	–	–
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	92.67	84.8	84.6	58.4	59.3	44.0	71.3	71.8	58.6	61.2	80.5	51.4
MCA	$98.1_{0.1}$	$95.5_{0.1}$	$96.0_{0.1}$	$92.7_{0.2}$	$84.9_{0.2}$	$84.6_{0.2}$	$59.7_{0.9}$	$59.8_{0.5}$	$44.0_{0.9}$	$74.9_{2.5}$	$73.3_{1.5}$	$61.6_{2.5}$	$61.2_{0.5}$	$79.7_{0.7}$	$51.9_{0.8}$
MCA	**98.2**	**95.5**	**96.0**	**92.8**	**85.0**	**84.9**	**61.2**	**60.6**	**45.5**	**77.9**	**75.1**	**64.3**	**61.9**	**81.1**	**52.3**

Table 1: Clustering results on five benchmark datasets. The best results are highlighted in bold.

is proportional to k'_I which reflects the neighborhood overlapping in the image domain, indicating that reducing the neighborhood imbalance (e.g., by setting a smaller number of neighbors k_I or filtering neighborhoods to reduce neighborhood imbalance) also reduces the expected risk of MCA.

Experiments and Analysis

In this section, experiments are conducted on five image benchmark datasets to validate the effectiveness of our proposed method.

Experimental Setup

Benchmarks and implementation details. We used the following five benchmark datasets in our experiment: STL10 (Coates, Ng, and Lee 2011), Cifar10 (Krizhevsky 2009), Cifar100-20 (Krizhevsky 2009), ImageNet-Dogs (Chang et al. 2017b) and Tiny-ImageNet (Le and Yang 2015).
Evaluation Metrics. We used three evaluation metrics to evaluate clustering results, including clustering Accuracy (ACC), Normalized Mutual Information (NMI) (McDaid, Greene, and Hurley 2011), and Adjusted Rand Index

(ARI) (Hubert and Arabie 1985). For these metrics, a higher value means better performance.

Comparisons with State-of-the-arts

Setup. We took the entire list of nouns in the WordNet dataset (Miller 1995) to form an initial semantic dataset for filtering which contains more than 82,000 nouns. To evaluate the effectiveness of our proposed method, we compare it with 22 state-of-the-art clustering methods on the five datasets, including $k$means (MacQueen 1967), SC (Zelnik-Manor 2005), NMF (Cai et al. 2009), JULE (Yang, Parikh, and Batra 2016), SAE (Ng et al. 2011), DAE (Vincent et al. 2010), AE (Bengio et al. 2006), VAE (Kingma and Welling 2013), 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 2019), PICA (Huang, Gong, and Zhu 2020), GCC (Zhong et al. 2021), CC (Li et al. 2021),TCL (Li et al. 2022),SCAN(Van Gansbeke et al. 2020), NNM (Dang et al. 2021) and SIC (Cai et al. 2023). We repeated the training five times independently on each dataset and reported their mean and standard deviation values.

Steps	Cifar10	ImageNet-Dogs
UF	91.4	65.3
UF+HF	91.9	71.8

Table 2: Ablation studies of semantic space construction (UF: Uniqueness-based filtering, HF: Hierarchy-based filtering.)

Loss Components				Result		
$\mathcal{L}_I$	$\mathcal{L}_{ia}$	$\mathcal{L}_{pa}$	$\mathcal{L}_{sa}$	ACC	NMI	ARI
✓				46.7±0.5	48.9 ±0.7	32.6 ±0.5
✓		✓	✓	66.0 ±3.4	67.7 ±2.5	50.7 ±4.1
✓	✓		✓	63.5 ±2.5	69.7 ±2.1	52.7 ±3.5
✓	✓	✓		53.1 ±3.8	55.4 ±2.8	37.4±2.4
✓	✓	✓	✓	**74.8 ±2.7**	**72.7 ±2.0**	**61.9 ±2.5**

Table 3: Ablation studies on ImageNet-Dogs.

Results. The comparison results with the state-of-the-art methods in terms of ACC, NMI, and ARI are presented in Tabel 1. From this table, we can observe that our method outperforms all other methods on five benchmark datasets. Especially, MCA improves ACC, NMI, and ARI by 2.8%, 1.3%, and 1.5% on Cifar100-20, and 6.6%, 3.3%, and 5.7% on ImageNet-Dogs, demonstrating that MCA better amend the incorrect alignments and thus achieve significant performance improvement. Especially for fine-grained images such as Imagenet-dogs whose categories are difficult to distinguish, our method can give useful clustering support information for images through the three alignment strategy, especially for fine-grained images like ImageNet-Dogs whose categories are difficult to distinguish.

Ablation Studies

Semantic space construction.

We first conduct an experiment on Cifar10 and ImageNet-Dogs to verify the effectiveness of our semantic space construction method and show the results in Table 2. The results demonstrate that **hierarchy-based filtering** can significantly enhance the semantic space compared to **uniqueness-based filtering**, especially when dealing with complex clusters that are challenging to distinguish, such as those in the ImageNet-Dogs dataset.
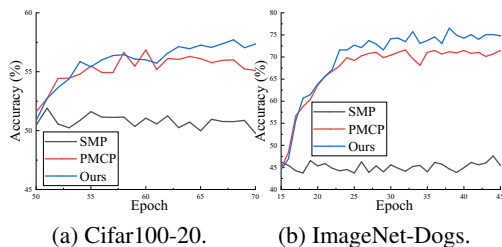

(a) Cifar100-20.   (b) ImageNet-Dogs.

Figure 3: The average accuracy of 10 runs of pseudo-labels with epochs on Cifar100-20 and ImageNet-Dogs evolves.
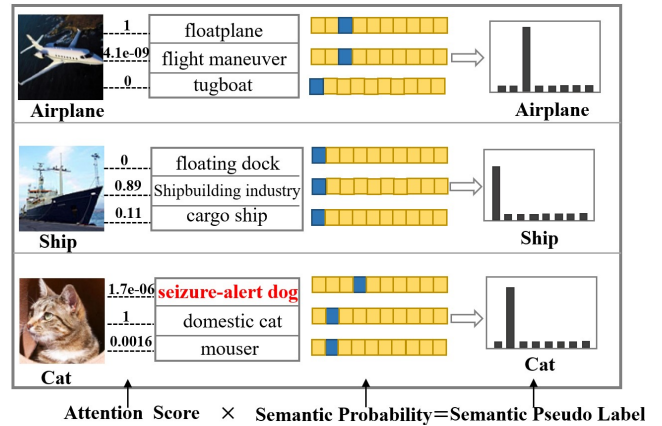

Figure 4: Example of pseudo-label generation in MCA. The words below (on the right side of) the images are ground-truth/ neighboring labels and the red color indicates irrelevant texts. The blue block in the semantic probability indicates the class the left word is assigned to (with the largest probability).

Loss components effectiveness. We perform an ablation analysis on ImageNet-Dogs to measure the importance of four loss components in our model, i.e., image consistency loss $\mathcal{L}_I$, instance-level alignment loss $\mathcal{L}_{ia}$, prototype-level alignment loss $\mathcal{L}_{pa}$ and semantic-level alignment $\mathcal{L}_{sa}$. The results are shown in Tabel 3, indicating that each of the four components plays an important role. The integration of three cross-modal alignment strategies significantly enhances the clustering performance, even obtaining 28.1%, 23.8%, and 29.3% performance gains when all three strategies are simultaneously used. Among the three cross-modal alignment strategies, introducing semantic-level alignment yields the most significant improvement in clustering performance, indicating that operating at the semantic level can effectively address incorrect alignments. These results confirm the effectiveness of our proposed cross-modal alignment methods.

Comparison of three pseudo-label generation methods. In our method, semantic-level cross-modal alignment can be considered self-training with cross-modal pseudo-labels. To verify the effectiveness of our method, we compared three pseudo-label generation methods implemented in our framework: 1) Single Modal Pseudo-labeling (**SMP**) that directly generates one-hot pseudo-labels via the argmax operation on the soft cluster assignments only from images, 2) Prototype Mapping based Cross-modal Pseudo-labeling (**PMCP**, is the adjusted center-based method in (Cai et al. 2023)) that generates pseudo-labels from the prototype level alignments in the original CLIP, and 3) **Ours** that generates pseudo-labels by simultaneously learning relationships among images and neighboring texts, while also updating the image and text embeddings. The comparison results on Cifar100-20 and ImageNet-Dogs are shown in Figure 3, demonstrating that our method significantly outperforms the other two methods on both datasets. These results indicate that learning adaptive relationships at the semantic level substantially improves the quality of pseudo-labels.
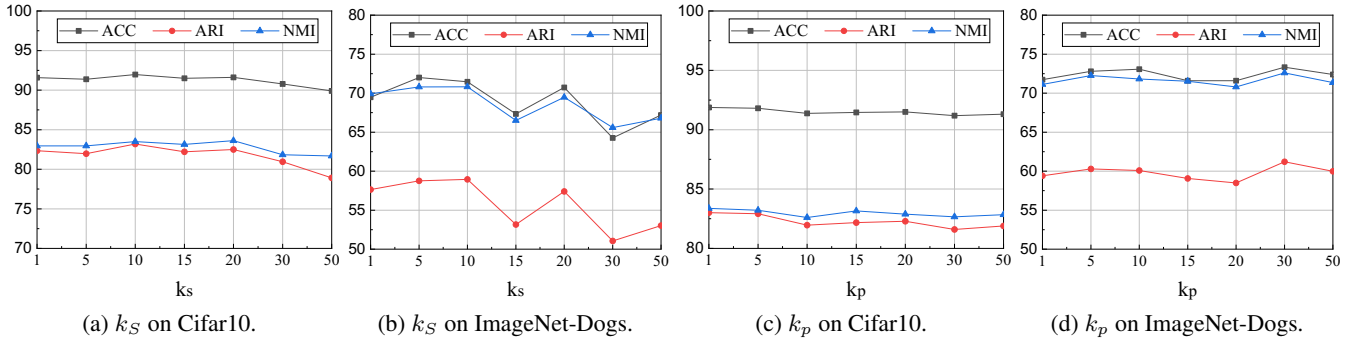
(a) k_S on Cifar10.   (b) k_S on ImageNet-Dogs.   (c) k_p on Cifar10.   (d) k_p on ImageNet-Dogs.

Figure 5: Sensitivity analysis of k_S and k_p.

(a) (η, λ_a) on ImageNet-Dogs.   (b) (η, λ_{pa}) on ImageNet-Dogs.   (c) (η, λ_{sa}) on ImageNet-Dogs.   (d) $(\lambda_{pa}, \lambda_{sa})$ on ImageNet-Dogs.
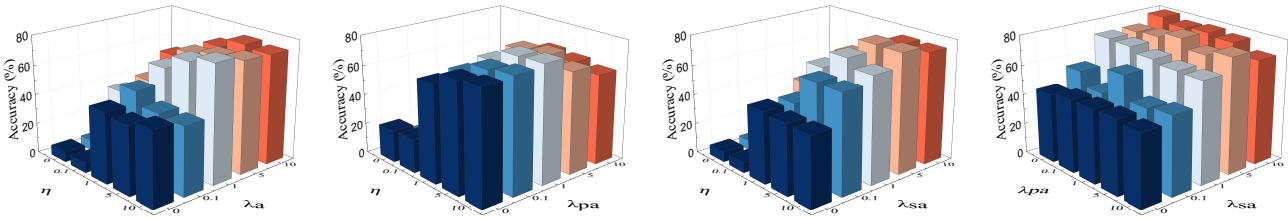
Figure 6: Sensitivity analysis of trade-off parameters η, λ_a, λ_{pa} and λ_{sa}.

Sensitivity Analysis

Figure 4 shows an example of our proposed pseudo-label generation method. We select an example from each of the three classes and three neighboring words for each image. Although these exist irrelevant neighboring words for an image, our method can identify irrelevant words and eliminate the affection of incorrect alignments in CLIP.

Sensitivity on neighborhood parameters k_S and k_p in cross-modal alignment. In our alignment method, k_S controls the number of neighboring texts in the instance-level and semantic-level alignments, and k_p controls the number of neighboring texts to recompute the semantic prototype in the prototype-level alignment. Figures 5a and 5b show that too large k_S causes performance degeneration due to the introduction of irrelevant texts. Figures 5c and 5d show that k_p does not change the performance too much.

Sensitivity on trade-off parameters η, λ_a, λ_{pa} and λ_{sa}. Figure 6 shows the sensitivity analysis of trade-off parameters η, λ_a, λ_{pa} and λ_{sa}. We can see that the performance of our method improves with increasing values of η, λ_{pa}, and λ_{sa}. Notably, our method appears to be more sensitive to changes in λ_{sa} than λ_{pa} on ImageNet-Dogs.

Compare with Zero-shot Learning

We compared MCA with two zero-shot learning methods, i.e., CLIP and MUST. Compared to CLIP, MUST improve CLIP by 3.3% on Caltech101 and 18% on UCF101. Compared to MUST, our method deteriorates MUST by -8.2%

on Caltech101 and -8.5% on UCF101. Although the results show that our method performs worse than CLIP and MUST, our method has the advantage of not requiring class names as input, which is necessary for zero-shot learning. This makes our approach more versatile and applicable to a broader range of real-world scenarios.

Conclusion

We have proposed a novel method to address the incorrect alignments in CLIP for image clustering. Our method includes the construction of a proper semantic space and a multi-level cross-modal alignment approach for aligning images and texts in downstream tasks at three levels. Theoretical results have shown interesting insights, and experimental results have demonstrated the superiority of our method. However, we acknowledge that our method may not be as cost-effective as SIC, as it involves three types of alignments. The proper setting of hierarchy levels also remains a challenge that needs further investigation. Additionally, our method exhibits lower performance compared to MUST, primarily due to the absence of class names. For future work, we will focus on enhancing the performance of our method and exploring new avenues for improvement. Leveraging the theoretical results to augment our method holds promise and will be a key research direction we pursue.

Acknowledgments

References

Bengio, Y.; Lamblin, P.; Popovici, D.; and Larochelle, H. 2006. Greedy layer-wise training of deep networks. In *Proceedings of neural information processing systems*, 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.

Cai, S.; Qiu, L.; Chen, X.; Zhang, Q.; and Chen, L. 2023. Semantic-Enhanced Image Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6869–6878.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, 132–149.

Chang, J.; Guo, Y.; Wang, L.; Meng, G.; Xiang, S.; and Pan, C. 2019. Deep Discriminative Clustering Analysis. *ArXiv*, abs/1905.01681.

Chang, J.; Wang, L.; Meng, G.; Xiang, S.; and Pan, C. 2017a. Deep Adaptive Image Clustering. In *In Proceedings of the IEEE international conference on computer vision*, 5880–5888.

Chang, J.; Wang, L.; Meng, G.; Xiang, S.; and Pan, C. 2017b. Deep Adaptive Image Clustering. In *Proceedings of 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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*.

Gao, Q.; Lian, H.; Wang, Q.; and Sun, G. 2020. Cross-modal subspace clustering via deep canonical correlation analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3938–3945.

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 the IEEE/CVF conference on computer vision and pattern recognition*, 8846–8855.

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

Ji, X.; Vedaldi, A.; and Henriques, J. F. 2019. Invariant Information Clustering for Unsupervised Image Classification and Segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 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.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint 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, J.; Savarese, S.; and Hoi, S. C. 2022. Masked unsupervised self-training for zero-shot image classification. *arXiv preprint arXiv:2206.02967*.

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, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; and Wang, H. 2020. 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.

Li, Y.; Yang, M.; Peng, D.; Li, T.; Huang, J.; and Peng, X. 2022. Twin contrastive learning for online clustering. *International Journal of Computer Vision*, 130(9): 2205–2221.

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.

Mao, Y.; Yan, X.; Guo, Q.; and Ye, Y. 2021. Deep mutual information maximin for cross-modal clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8893–8901.

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.; et al. 2011. Sparse autoencoder. *CS294A Lecture notes*, 72(2011): 1–19.

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 *Proceedings of International Conference of Learning Representation*, 1–20.

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 European conference on computer vision*, 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).

Wu, J.; Long, K.; Wang, F.; Qian, C.; Li, C.; Lin, Z.; and Zha, H. 2019. Deep Comprehensive Correlation Mining for Image Clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8149–8158.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised Deep Embedding for Clustering Analysis. In *Proceedings of International conference on machine learning*, 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 international conference on machine learning*, volume 70, 3861–3870.

Yang, J.; Parikh, D.; and Batra, D. 2016. Joint Unsupervised Learning of Deep Representations and Image Clusters. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5147–5156.

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.