

# On the Provable Importance of Gradients for Autonomous Language-Assisted Image Clustering

Bo Peng, Jie Lu, Guangquan Zhang, Zhen Fang\*  
University of Technology Sydney  
Sydney, Australia

## Abstract

*This paper investigates the recently emerged problem of Language-assisted Image Clustering (LaIC), where textual semantics are leveraged to improve the discriminability of visual representations to facilitate image clustering. Due to the unavailability of true class names, one of core challenges of LaIC lies in how to filter positive nouns, i.e., those semantically close to the images of interest, from unlabeled wild corpus data. Existing filtering strategies are predominantly based on the off-the-shelf feature space learned by CLIP; however, despite being intuitive, these strategies lack a rigorous theoretical foundation. To fill this gap, we propose a novel gradient-based framework, termed as GradNorm, which is theoretically guaranteed and shows strong empirical performance. In particular, we measure the positiveness of each noun based on the magnitude of gradients back-propagated from the cross-entropy between the predicted target distribution and the softmax output. Theoretically, we provide a rigorous error bound to quantify the separability of positive nouns by GradNorm and prove that GradNorm naturally subsumes existing filtering strategies as extremely special cases of itself. Empirically, extensive experiments show that GradNorm achieves the state-of-the-art clustering performance on various benchmarks.*

## 1. Introduction

As a fundamental problem in pattern recognition and machine learning, image clustering [36] seeks to separate a set of unlabeled images into multiple groups such that images in the same group are semantically similar to each other. Due to its ability to reveal the inherent semantic structure underlying the data without requiring laborious and trivial data labeling work, clustering has been shown to benefit downstream tasks [4, 5, 27, 43–46, 69, 72, 74, 75, 77–79] in computer vision. Despite increasing attention, the vast majority of strategies [18, 26, 34, 39, 61, 64, 66] to image clus-

tering reply on purely visual supervision signals and therefore inherit limitations especially when images of interest are visually similar to but semantically different from each other.

This paper delves into a new landscape for image clustering by departing from the classic single-model toward a multi-modal regime. In the visual domain, (deep) clustering methods usually learn discriminative representations from distributional priors [22, 39, 49], pseudo-labels [2, 3, 17, 62], neighborhood consistency [10, 56, 70] and augmentation invariance [12, 29, 31], which, however, can not be directly transferred into the vision-language regime due to the heterogeneous relation between visual and textual data. While the advanced vision-language pre-training schemes, e.g., CLIP [50], have emerged as promising alternatives for visual representation learning by mapping textual and visual inputs into a unified representation space, harnessing the power of texts to facilitate image clustering is still non-trivial due to the *unavailability of class name priors*.

To address this challenge, the mainstream solutions [1, 32] are to select positive nouns, i.e., those who best describe images of interest, from unlabeled lexical databases in the wild<sup>1</sup> (e.g., WordNet [38]) for the textual pseudo-labeling of each image. Despite recent empirical successes, the *separability of positive nouns* remains theoretically underexplored, with no prior work providing a rigorous formalization or provable error bounds. Our work thus complements existing works by filling in the critical blank. In this paper, we design a simple yet effective framework that provides a provable guarantee for Language-assisted Image Clustering (LaIC) from a novel perspective of gradient.

Methodologically, our proposed method GradNorm begins by learning a single-layer self-supervised classifier using CLIP features extracted from pseudo-labeled images. Leveraging the alignment between the CLIP image and text feature spaces [7], we extend the learned classifier to handle text features as well. Subsequently, we employ CLIP features of unlabeled wild texts as input to compute the gra-

<sup>1</sup>Generally, “in-the-wild” data are those that can be collected almost for free upon deploying machine learning models in the open world.

\*Correspondence Author

dients of the classifier back-propagated based on the cross-entropy between the softmax output and the predicted target distribution. In this process, we consider unlabeled wild nouns as positive samples if the magnitude of the corresponding gradients falls below an adaptive threshold.

Theoretically, we justify GradNorm in Theorem 1 and Section 4. Our theoretical insights are twofold. First, we derive a rigorous upper bound on the error rate for separating positive nouns from unlabeled wild data. This upper bound is proportional to the optimal risk, which can approach zero in practice especially when the size of the pre-trained CLIP model is sufficiently large. Second, our analysis establishes a unified framework for existing filtering strategies [1, 32] by demonstrating that, despite their apparent differences in motivation and methodology, they can be interpreted as de-generated cases of GradNorm.

Extensive experiments on multiple benchmarks demonstrate the empirical effectiveness of our proposed GradNorm method. For example, GradNorm achieves 60.6% ACC and 81.2% ACC on CIFAR-20 and ImageNet-Dog datasets, respectively, outperforming the latest TAC [32] by 4.8% and 6.1%. Additionally, on three more challenging datasets (DTD, UCF-101, and ImageNet-1K), our method surpasses TAC [32] by an average of 3.2%, 1.7%, and 2.4% in terms of ACC, NMI, and ARI, respectively.

## 2. Related Work

### 2.1. Deep Image Clustering

The popularity of deep image clustering can be attributed to the fact that distributional assumptions in classic clustering methods, e.g., compactness [14], connectivity [42, 58], sparsity [73, 76] and low rankness [33], can not be necessarily conformed by high-dimensional structural RGB images. To exploit the powerful representative ability of deep neural networks in an unsupervised manner, the earliest attempts seeks self-supervision signals by considering image reconstruction [15, 47, 61], probabilistic modeling [22, 39, 49] and mutual information maximization [16, 20] as proxy tasks. Despite remarkable progresses, the learned representations may not be discriminative enough to capture the semantic similarity between images. More recently, the advance in self-supervised representation learning have led to major breakthroughs in deep image clustering. On the one hand, IDFD [53] proposes to perform both instance discrimination and feature de-correlation while MICE [54] propose a unified latent mixture model based on contrastive learning to tackle the clustering task. On the other hand, CC [29] and its followers TCC [31] perform contrastive learning at both instance and cluster levels. Different from above methods, ProPos [18] performs non-contrastive learning on the instance level and contrastive learning on the cluster level, which results in enjoying the strengths of both worlds.

### 2.2. Vision-language Models

Leveraging large-scale pre-trained vision-language models (VLMs) has emerged as a remarkably effective paradigm for multi-modal downstream tasks. Regarding the type of architectures, existing VLMs can be divided into two categories: 1) single-stream models like VisualBERT [28] and ViLT [23] feed the concatenated text and visual features into a single transformer-based encoder; 2) dual-stream models such as CLIP [50], ALIGN [21], and FILIP [65] use separate encoders for text and image and optimize with contrastive objectives to align semantically similar features in different modalities. In particular, CLIP enjoys popularity due to its simplicity and strong performance. CLIP-like models inspire numerous follow-up works [30, 68, 71] that aim to improve data efficiency and better adaptation to downstream tasks. This paper uses CLIP as the pre-trained model, but our method can be generally applicable to contrastive models that promote vision-language alignment.

### 2.3. Language-assisted Image Clustering

The core of LaIC lies in how to leverage texture semantics as the supervision signal to guide clustering in the visual domain. The seminar work called SIC [1] uses textual semantics to enhance image pseudo-labeling, followed by performing image clustering with consistency learning in both image space and semantic space. Note that, SIC essentially pulls image embeddings closer to embeddings in semantic space, while ignoring the improvement of text semantic embeddings. Differently, TAC [32] focuses on leveraging textual semantics to enhance the feature discriminability by either simply concentrating textual and visual features or its proposed cross-modal mutual distillation strategy. Despite their variety in the usage of texture semantics for image clustering, both SIC and TAC requires filter positive semantics from unlabeled wild textual data due to the lack of true class names. However, to the best of our knowledge, a formalized understanding regarding the separation of positive semantics is currently lacking for this field, which directly motivates our work.

## 3. Proposed Framework: GradNorm

### 3.1. Preliminary: Zero-shot Classification

Let  $\mathcal{X}$  and  $\mathcal{T}$  be the visual and textual input space respectively, CLIP-based models adopt a simple dual-stream architecture with one text encoder  $f_{\mathcal{T}}$  and one image encoder  $f_{\mathcal{X}}$  to map inputs of two modalities into a uni-modal hyperspherical feature space  $\mathcal{Z} \triangleq \{\mathbf{z} \in \mathbb{R}^d \mid \|\mathbf{z}\|_2 = 1\}$ . Considering an image classification task with the known classes  $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$ , CLIP-based models make class prediction for any input image  $\mathbf{x} \in \mathcal{X}$  by computing the following

$$\arg \max_{i=1, \dots, K} \frac{\exp[\tau f_{\mathcal{X}}(\mathbf{x})^\top f_{\mathcal{T}}(\Delta(\mathbf{c}_i))]}{\sum_{j=1}^K \exp[\tau f_{\mathcal{X}}(\mathbf{x})^\top f_{\mathcal{T}}(\Delta(\mathbf{c}_j))]}, \quad (1)$$

where  $\tau > 0$  is a temperature hyper-parameter,  $\Delta(\mathbf{c}_i) \in \mathcal{T}$  with  $\Delta(\cdot)$  as the prompt template for the input class name.

### 3.2. Leveraging Unlabeled Textual Data in the Wild

Despite remarkable effectiveness [50] and provable guarantees [6], the zero-shot paradigm in Eq. (1) suffers from the reliance on the prior knowledge of true class names, therefore inapplicable to the task of image clustering since we have access to only the number of ground-truth classes  $K$ .

In this paper, we address this challenge by leveraging unlabeled ‘‘in-the-wild’’ textual data which can be collected almost for free in the open world. However, it is important to note that wild textual data inevitably contains a mixture of *positive*<sup>2</sup> and *negative semantics* regarding to the image dataset of interest. In view of this, we propose to use the Huber contamination model [19] to model the marginal distribution of the wild textual data as follows:

**Definition 1** (Wild Data Distribution). *Let  $\mathbb{P}_{pos}$  and  $\mathbb{P}_{neg}$  be the distributions of positive and negative textual data defined over  $\mathcal{T}$ , respectively. According to the Huber contamination model [19], we can model the unlabeled textual data distribution  $\mathbb{P}_{wild}$  as follows:*

$$\mathbb{P}_{wild} \triangleq \pi \cdot \mathbb{P}_{pos} + (1 - \pi) \cdot \mathbb{P}_{neg}, \quad (2)$$

where  $\pi \in (0, 1]$  is typically unknown in practice.

**Definition 2** (Empirical Wild Dataset). *An empirical wild textual dataset  $\mathcal{D}_{\mathcal{T}}$  is sampled independently and identically distributed (i.i.d.) from the wild data distribution  $\mathbb{P}_{wild}$ .*

Following prior works [1, 32], we simulate the wild dataset  $\mathcal{D}_{\mathcal{T}}$  by resorting to the off-the-shelf WordNet [38]. In particular, let  $\{\tilde{\mathbf{c}}_1, \dots, \tilde{\mathbf{c}}_M\}$  be a pre-defined subset of nouns from WordNet, we can write  $\mathcal{D}_{\mathcal{T}} = \{\tilde{\mathbf{t}}_i = \Delta(\tilde{\mathbf{c}}_i)\}_{i=1}^M$ .

**Remark 1.** *While wild textual data can be available in abundant without requiring human annotations, harnessing such data is non-trivial due to the lack of clear membership (either positive or negative) for textual data in  $\mathcal{D}_{\mathcal{T}}$ . Therefore, we aim to devise an automated strategy that estimates the membership for samples within the unlabeled textual data, therefore enabling the assistance of language for image clustering. In what follows, we describe these two stages in Section 3.3 and Section 3.4 respectively.*

<sup>2</sup>Examples of positive semantics can be synonyms of the name of objects in given images or those of adjectives that describe objects in given images.

### 3.3. Filtering Candidate Positive Semantics

**Overview.** To separate candidate positive semantics from the wild dataset  $\mathcal{D}_{\mathcal{T}}$ , we employ a level-set estimation based on the gradient information. The gradients are estimated from a classifier trained on the pseudo-labeled images. We describe the procedure formally below.

#### 3.3.1. Classifier Pre-training

To realize the idea, let  $\mathcal{D}_{\mathcal{X}} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$  denotes the image dataset of interest, we begin with extracting features from CLIP-based models for images in the dataset  $\mathcal{D}_{\mathcal{X}}$  to have  $\mathbf{E} = (\mathbf{e}_1, \dots, \mathbf{e}_N) \in \mathbb{R}^{d \times N}$  where  $\mathbf{e}_i = f_{\mathcal{X}}(\mathbf{x}_i) \in \mathcal{Z}$  for each  $i \in [N] := \{1, \dots, N\}$ . By performing a classical clustering algorithm, e.g.,  $k$ -means, on the image feature matrix  $\mathbf{E}$  to grouping given images into  $C$  clusters, we can produce pseudo-label  $y_i \in \mathcal{Y} \triangleq [C]$  for each image  $\mathbf{x}_i \in \mathcal{D}_{\mathcal{X}}$  to learn a single-layer classifier  $h(\cdot; \mathbf{W}) : \mathcal{Z} \rightarrow \mathbb{R}^C$  parameterized by  $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_C) \in \mathbb{R}^{d \times C}$  with the following empirical risk minimization (ERM):

$$\mathbf{W}^* = \arg \min_{\mathbf{W} \in \mathcal{W}} \frac{1}{N} \sum_{i=1}^N \ell(h(\mathbf{e}_i; \mathbf{W}), y_i), \quad (3)$$

where  $\mathcal{W}$  is the parameter space and  $\ell(h(\mathbf{e}_i; \mathbf{W}), y_i)$  is the cross-entropy between the softmax output  $h(\mathbf{x}_i; \mathbf{W}) = \text{softmax}(\tau \cdot \mathbf{e}_i^\top \mathbf{W})$  and the pseudo-target distribution, i.e.,

$$\ell(h(\mathbf{e}_i; \mathbf{W}), y_i) \triangleq -\log \frac{\exp(\tau \mathbf{e}_i^\top \mathbf{w}_{y_i})}{\sum_{k \in [C]} \exp(\tau \mathbf{e}_i^\top \mathbf{w}_k)}. \quad (4)$$

#### 3.3.2. Membership Estimation via Gradient Norm

Key to this step, we perform a scoring procedure to measure the positiveness of each text in the wild dataset  $\mathcal{D}_{\mathcal{T}}$  to  $\mathcal{D}_{\mathcal{X}}$ , the image dataset of interest. To formulate the score function  $S$ , we forward the feature of each text in the wild dataset  $\mathcal{D}_{\mathcal{T}}$  into the learned classifier  $h(\cdot; \mathbf{W}^*)$  to calculate the gradients w.r.t. the classifier parameters  $\mathbf{W}^*$  by back-propagating the cross entropy between the softmax output and the predicted target distribution. In particular, let  $\tilde{\mathbf{R}} = (\tilde{\mathbf{r}}_1, \dots, \tilde{\mathbf{r}}_M) \in \mathbb{R}^{d \times M}$  as the textual feature matrix for the wild dataset  $\mathcal{D}_{\mathcal{T}}$  where  $\tilde{\mathbf{r}}_i = f_{\mathcal{T}}(\tilde{\mathbf{t}}_i) \in \mathcal{Z}$  for each  $\tilde{\mathbf{t}}_i \in \mathcal{D}_{\mathcal{T}}$ , we define the gradient matrix  $\mathbf{G}$  as follows:

$$\mathbf{G} = \begin{bmatrix} \partial \ell(h(\tilde{\mathbf{r}}_1; \mathbf{W}^*), \tilde{y}_1) / \partial \mathbf{W}^* \\ \vdots \\ \partial \ell(h(\tilde{\mathbf{r}}_M; \mathbf{W}^*), \tilde{y}_M) / \partial \mathbf{W}^* \end{bmatrix}, \quad (5)$$

where  $\tilde{y}_i = \arg \min_{k \in [C]} \ell(h(\tilde{\mathbf{r}}_i; \mathbf{W}^*), k)$ . To assign the membership with  $\tilde{\mathbf{t}}_i \in \mathcal{D}_{\mathcal{T}}$ , we define the estimation score

$S$  as follows<sup>3</sup>:

$$\begin{aligned} S(\tilde{\mathbf{t}}_i) &= \left\| \frac{\partial \ell(h(\tilde{\mathbf{r}}_i; \mathbf{W}^*), \tilde{y}_i)}{\partial \mathbf{W}^*} \right\|_F^2 \\ &= \tau^2 \cdot \left( \sum_{k \in [C]} \tilde{\pi}_{ik}^2 + 1 - 2 \max_{j \in [C]} \tilde{\pi}_{ij} \right), \end{aligned} \quad (6)$$

where  $\|\cdot\|_F$  denotes the Frobenius norm and

$$\tilde{\pi}_{ij} = \frac{\exp(\tau \cdot \tilde{\mathbf{r}}_i^\top \mathbf{w}_j^*)}{\sum_{k \in [C]} \exp(\tau \cdot \tilde{\mathbf{r}}_i^\top \mathbf{w}_k^*)}, \quad \forall j \in [C]. \quad (7)$$

Finally, we can arrive at the (potentially noisy) set of candidate positive text semantics as follows:

$$\hat{\mathcal{P}}_{\mathcal{T}}(k) \triangleq \left\{ \tilde{\mathbf{t}}_i \in \mathcal{D}_{\mathcal{T}} : S(\tilde{\mathbf{t}}_i) \leq T_k \text{ and } k = \arg \max_{j \in [C]} \tilde{\pi}_{ij} \right\}, \quad (8)$$

where  $T_k$  denotes the  $\beta$ -th smallest score of text semantics in the set  $\{\tilde{\mathbf{t}}_i \in \mathcal{D}_{\mathcal{T}} : k = \arg \max_{j \in [C]} \tilde{\pi}_{ij}\}$ . In the following, our main theorem formally quantifies the separability of truly positive text semantics from the wild dataset  $\mathcal{D}_{\mathcal{T}}$  by leveraging the filtering strategy in Eq. (8).

**Theorem 1.**<sup>4</sup> *Let us define the ground-truth set of truly positive semantics from the wild data as*

$$\mathcal{P}_{\mathcal{T}}(k) = \left\{ \tilde{\mathbf{t}}_i \in \mathcal{D}_{\mathcal{T}} : \tilde{\mathbf{t}}_i \sim \mathbb{P}_{pos} \text{ and } k = \arg \max_{j \in [C]} \tilde{\pi}_{ij} \right\}$$

and  $|\mathcal{P}_{\mathcal{T}}(k)| = B_k$ . Under mild assumptions (cf. Appendix. 3), i.e., the loss function  $\ell$  is  $\gamma$ -smooth and the parameter space  $\mathcal{W}$  is bounded, with the probability at least 0.97, we have the following:

$$\begin{aligned} Err_{pos}(k) &\triangleq \frac{|\{\tilde{\mathbf{t}}_i \in \mathcal{P}_{\mathcal{T}}(k) : S(\tilde{\mathbf{t}}_i) > T_k\}|}{O_k} \\ &\leq \frac{2\gamma}{T_k} \left[ \min_{\mathbf{W} \in \mathcal{W}} \Omega(\mathbf{W}) + O\left(\sqrt{\frac{1}{B_k}}\right) + O\left(\sqrt{\frac{1}{N}}\right) \right], \end{aligned}$$

where  $\Omega(\mathbf{W}) = \mathbb{E}_{(\mathbf{z}, y) \sim \mathbb{P}_{\mathcal{Z}\mathcal{Y}}} \ell(h(\mathbf{z}; \mathbf{W}), y)$  is the expected risk and we use  $O(\cdot)$  to hide universal constant factors.

**Remark 2.** *Theorem 1 states that, under mild assumptions,  $ERR_{pos}(k)$  is upper-bounded. In particular, if the following two regulatory conditions hold: 1) the size of the image data  $N$  and that of the wild textual data  $B_k$  are sufficiently large; 2) the minimal expected risk  $\min_{\mathbf{W} \in \mathcal{W}} \Omega(\mathbf{W})$  is sufficiently small, then the upper bound is also small.*

<sup>3</sup>We provide detailed deviation of the second step in the appendix

<sup>4</sup>Due to space limitation, we defer detailed proofs in the appendix.

---

### Algorithm 1 Pipeline of GradNorm

---

**Input:** Image features  $\{\mathbf{e}_i\}_{i=1}^N$ , Text features  $\{\tilde{\mathbf{r}}_i\}_{i=1}^M$ , Randomly initialized parameters  $\mathbf{W}$

▷ *Stage 1: Filtering Candidate Positive Semantics*

- 1: Apply  $k$ -means on image features  $\{\mathbf{e}_i\}_{i=1}^N$  to obtain pseudo-labels  $\{y_i \in [C]\}_{i=1}^N$
- 2: Obtain  $\mathbf{W}^*$  by performing ERM in Eq. (3)
- 3: Compute  $S(\tilde{\mathbf{t}}_i)$  in Eq. (6) to obtain  $\hat{\mathcal{P}}_{\mathcal{T}}(k)$  via Eq. (8)

- 4: Get candidate positive semantics  $\hat{\mathcal{D}}_{\mathcal{T}}^{pos} = \bigcup_{k=1}^C \hat{\mathcal{P}}_{\mathcal{T}}(k)$ .

▷ *Stage 2: Clustering with Candidate Positive Semantics*

- 5: Compute  $\mathbf{v}_i$  for each  $\mathbf{e}_i$  via Eq. (9)
  - 6: Apply  $k$ -means on the concatenated image-text features  $\{[\mathbf{e}_i; \mathbf{v}_i]\}_{i=1}^N$  to obtain final cluster assignment
- 

### 3.4. Clustering with Candidate Positive Semantics

After extracting the candidate positive semantics set  $\hat{\mathcal{D}}_{\mathcal{T}}^{pos} = \bigcup_{k=1}^C \hat{\mathcal{P}}_{\mathcal{T}}(k)$  from the wild textual dataset  $\mathcal{D}_{\mathcal{T}}$ , it is essential to design an effective collaboration mechanism between text semantics and image semantics for clustering. Given that the primary contribution of this paper is to reliably select positive semantics from unlabeled wild textual data, we adopt the same post-hoc collaboration strategy as [32].

In particular, for each image  $\mathbf{x}_i$ , we build the corresponding text counterpart  $\mathbf{v}_i$  by resorting to deep set representations [51, 67], i.e.,

$$\mathbf{v}_i = \sum_{\tilde{\mathbf{t}}_j \in \hat{\mathcal{D}}_{\mathcal{T}}^{pos}} \left( \frac{\exp(\mathbf{e}_i^\top \tilde{\mathbf{r}}_j / \kappa)}{\sum_{\tilde{\mathbf{t}}_k \in \hat{\mathcal{D}}_{\mathcal{T}}^{pos}} \exp(\mathbf{e}_i^\top \tilde{\mathbf{r}}_k / \kappa)} \cdot \tilde{\mathbf{r}}_j \right), \quad (9)$$

where  $\kappa > 0$  is a temperature hyper-parameter.

Finally, we compute the cluster assignment for the image dataset  $\mathcal{D}_{\mathcal{X}}$  by applying  $k$ -means on the concatenated image-text features  $\{[\mathbf{e}_i; \mathbf{v}_i] \in \mathbb{R}^{2d}\}_{i=1}^N$ . For clarity, we summarize the details of GradNorm in Algorithm 1.

## 4. Discussions

In this section, we discuss the theoretical connection between our method between prior works [1, 32] by showing that the latter can be explained as extremely special cases of the former though they indeed seem to be quite distinct regarding their proposed filtering strategies. In particular, our theoretical analysis is motivated by SeCu [48] to consider training the classifier  $h(\cdot; \mathbf{W})$  with the following objective:

$$\hat{\ell}(h(\mathbf{e}_i; \mathbf{W}), y_i) \triangleq -\log \frac{\exp(\tau \mathbf{e}_i^\top \mathbf{w}_{y_i})}{\exp(\tau \mathbf{e}_i^\top \mathbf{w}_{y_i}) + \sum_{k \neq y_i} \exp(\tau \mathbf{e}_i^\top \hat{\mathbf{w}}_k)}, \quad (10)$$

where  $\hat{\mathbf{w}} = sg(\mathbf{w})$  with  $sg(\cdot)$  as the stop-gradient operator.

**Remark 3.** Clearly,  $\hat{\ell}$  in Eq. (10) differs from the standard cross entropy in Eq. (4) in that each weight vector  $\mathbf{w}_k$  is only updated by image features whose pseudo label  $y_i = k$ . While it has been shown in SeCu [48] that  $\hat{\ell}$  in Eq. (10) can be more stable than the standard cross entropy when the size of training batch is so small that the weight vector  $\mathbf{w}_k$  is only updated by image features whose pseudo label  $y_i \neq k$ , we note that, since the memory complexity of the classifier  $h(\cdot; \mathbf{W})$  is only  $O(d \cdot C)$ , training with large batches (e.g., 2048) can be applicable in this paper.

**Theorem 2** ([48]). Let  $\mathbf{W}^* = (\mathbf{w}_1^*, \dots, \mathbf{w}_C^*)$  be the empirical risk minimizer of the loss function in Eq. (4) over the dataset  $\{(\mathbf{e}_i, y_i)\}_{i=1}^N$ . If we fix  $\|\mathbf{w}_k\|_2 = 1$  for any  $k \in [C]$ , we then arrive at the closed form of  $\mathbf{W}^*$  given by:

$$\mathbf{w}_j^* = \Lambda \left( \frac{\sum_{i:y_i=j} (1 - \pi_{ij}) \mathbf{e}_i}{\sum_{i:y_i=j} (1 - \pi_{ij})} \right), \quad (11)$$

where the operator  $\Lambda(\cdot)$  denotes the  $L_2$ -normalizer and

$$\pi_{ij} = \frac{\exp(\tau \mathbf{e}_i^\top \mathbf{w}_j^*)}{\exp(\tau \mathbf{e}_i^\top \mathbf{w}_j^*) + \sum_{k \neq y_i} \exp(\tau \mathbf{e}_i^\top \hat{\mathbf{w}}_k^*)}.$$

#### 4.1. Connection to TAC [32]

In the extremely special case where  $\tau \rightarrow 0$ , we have  $\pi_{ij} \rightarrow 1/C$  to approximate  $\mathbf{w}_j^*$  in Eq. (11) as the center of image features that belongs to  $k$ -th cluster:

$$\mathbf{w}_j^* \rightarrow \Lambda \left( \frac{\sum_{i:y_i=j} \mathbf{e}_i}{\sum_i \mathbb{I}(y_i = j)} \right) \text{ as } \tau \rightarrow 0, \quad (12)$$

where  $\mathbb{I}(\cdot)$  is the indicator function. In this way, we can arrive at the same maximum softmax probability (MSP)-based filtering score used in TAC [32] as a special case of our proposed score in Eq. (6), i.e.,

$$\begin{aligned} \hat{S}(\tilde{\mathbf{t}}_i) &= \left\| \frac{\partial \hat{\ell}(h(\tilde{\mathbf{r}}_i; \mathbf{W}^*), \tilde{y}_i)}{\partial \mathbf{W}^*} \right\|_F^2 \\ &= \left\| \frac{\partial \ell(h(\tilde{\mathbf{r}}_i; \mathbf{W}^*), \tilde{y}_i)}{\partial \mathbf{w}_{\tilde{y}_i}^*} \right\|_2^2 \propto \left( 1 - \max_{j \in [C]} \tilde{\pi}_{ij} \right)^2, \end{aligned} \quad (13)$$

so that  $\hat{S}(\tilde{\mathbf{t}}_i) \leq T_k \Leftrightarrow \max_{j \in [C]} \tilde{\pi}_{ij} \geq T_k'$  as  $\max_{j \in [C]} \tilde{\pi}_{ij} \leq 1$ .

**Remark 4.** It is important to note that the effectiveness of MSP-based score function  $\hat{S}$  in Eq. (13) can be challenged by the notorious overconfidence phenomenon [40] where neural networks tend to produce overconfident predictions, i.e., abnormally high softmax confidences, even when the inputs are far away from the training data.

#### 4.2. Connection to SIC [1]

**Assumption 1** (Self-normalization [54, 60]). An unnormalized classifier  $h(\cdot, \mathbf{W})$  is self-normalized, i.e., for any possible input  $z \in \mathcal{Z}$ ,  $\sum_{k \in [C]} \exp(\mathbf{z}^\top \mathbf{w}_k / \tau) = \text{const}$ , so that

$$\tilde{\pi}_{ij} = \frac{\exp(\tau \tilde{\mathbf{r}}_i^\top \mathbf{w}_j^*)}{\sum_{k \in [C]} \exp(\tau \tilde{\mathbf{r}}_i^\top \mathbf{w}_k^*)} \propto \exp(\tau \tilde{\mathbf{r}}_i^\top \mathbf{w}_j^*), \forall j \in [C].$$

In the extremely special case where  $\tau \rightarrow 0$ , if Assumption 1 holds for the classifier  $h(\cdot, \mathbf{W}^*)$  given by Eq. (12), we have:

$$\arg \max_{j \in [C]} \tilde{\pi}_{ij} = \arg \max_{j \in [C]} \tilde{\mathbf{r}}_i^\top \mathbf{w}_j^*. \quad (14)$$

Combining Eq. (14) and Eq. (13), we can arrive at the same cosine similarity-based scoring function used in SIC [1] as a special case of our proposed score in Eq. (6), i.e.,

$$\hat{S}(\tilde{\mathbf{t}}_i) \leq T_k \Leftrightarrow \max_{j \in [C]} \tilde{\pi}_{ij} \geq T_k' \Leftrightarrow \max_{j \in [C]} \tilde{\mathbf{r}}_i^\top \mathbf{w}_j^* \geq T_k''. \quad (15)$$

### 5. Experiments

#### 5.1. Experimental Setups

##### 5.1.1. Datasets

We evaluate the effectiveness of GradNorm by conducting experiments on 1) five widely-used datasets: STL-10 [9], CIFAR-10 [25], CIFAR-20 [25], ImageNet-10 [3], and ImageNet-Dogs [3]; 2) three more complex datasets with larger cluster numbers: DTD [8], UCF-101 [52], and ImageNet-1K [11]. Following prior works [1, 32], we filter candidate positive semantics based on the train split of each image dataset, followed by evaluate the clustering performance on the test split of each image dataset. To keep the main content concise, We summarize the details of these datasets in the appendix.

##### 5.1.2. Implementation Details

For a fair comparison with previous works [1, 32, 50], we, unless explicitly stated, adopt the pre-trained CLIP model with ViT-B/32 [13] and Transformer [57] as default image and text backbones, respectively. For nouns from WordNet [38], we assemble them with prompts like ‘‘A photo of [CLASS]’’ before feeding them into the Transformer. To find semantics of appropriate granularity given a  $N$ -sized image dataset, we, similar to TAC [32], set  $C = N/600$  for datasets with an average cluster size larger than 600 and  $C = 3K$  otherwise. We fix  $\tau = 1/0.08$ ,  $\kappa = 0.006$  and  $\beta = 5$  for all datasets. In most cases, we train the classifier  $h$  by the Adam [24] optimizer for 30 epochs with learning rate as  $1e - 3$  and batch size as 2048. The only exception is that on UCF-101 and ImageNet-1K, where, the classifier  $h$  is trained for 100 epochs with batch size as 8192. All experiments are conducted on a single Nvidia A100 GPU.

Table 1. Clustering performance (%) on five widely used image clustering datasets. The best results are highlighted in bold.

Dataset	STL-10			CIFAR-10			CIFAR-20			ImageNet-10			ImageNet-Dogs		
Metrics	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
CLIP (zero-shot)	93.9	97.1	93.7	80.7	90.0	79.3	55.3	58.3	39.8	95.8	97.6	94.9	73.5	72.8	58.2
JULE (CVPR16) [63]	18.2	27.7	16.4	19.2	27.2	13.8	10.3	13.7	3.3	17.5	30.0	13.8	5.4	13.8	2.8
DEC (ICML16) [61]	27.6	35.9	18.6	25.7	30.1	16.1	13.6	18.5	5.0	28.2	38.1	20.3	12.2	19.5	7.9
DAC (ICCV17) [3]	36.6	47.0	25.7	39.6	52.2	30.6	18.5	23.8	8.8	39.4	52.7	30.2	21.9	27.5	11.1
DCCM (ICCV19) [59]	37.6	48.2	26.2	49.6	62.3	40.8	28.5	32.7	17.3	60.8	71.0	55.5	32.1	38.3	18.2
IIC (ICCV19) [20]	49.6	59.6	39.7	51.3	61.7	41.1	22.5	25.7	11.7	—	—	—	—	—	—
PICA (CVPR20) [17]	61.1	71.3	53.1	59.1	69.6	51.2	31.0	33.7	17.1	80.2	87.0	76.1	35.2	35.3	20.1
CC (AAAI21) [29]	76.4	85.0	72.6	70.5	79.0	63.7	43.1	42.9	26.6	85.9	89.3	82.2	44.5	42.9	27.4
IDFD (ICLR20) [53]	64.3	75.6	57.5	71.1	81.5	66.3	42.6	42.5	26.4	89.8	95.4	90.1	54.6	59.1	41.3
SCAN (ECCV20) [56]	69.8	80.9	64.6	79.7	88.3	77.2	48.6	50.7	33.3	—	—	—	61.2	59.3	45.7
MiCE (ICLR20) [54]	63.5	75.2	57.5	73.7	83.5	69.8	43.6	44.0	28.0	—	—	—	42.3	43.9	28.6
GCC (ICCV21) [70]	68.4	78.8	63.1	76.4	85.6	72.8	47.2	47.2	30.5	84.2	90.1	82.2	49.0	52.6	36.2
NNM (CVPR21) [10]	66.3	76.8	59.6	73.7	83.7	69.4	48.0	45.9	30.2	—	—	—	60.4	58.6	44.9
CRLC (ICCV21) [12]	72.9	81.8	62.8	67.9	79.9	63.4	41.6	42.5	26.3	83.1	85.4	75.9	48.4	46.1	59.7
TCC (NeurIPS21) [51]	73.2	81.4	68.9	79.0	90.6	73.3	47.9	49.1	31.2	84.8	89.7	82.5	55.4	59.5	41.7
TCL (IJCV22) [31]	79.9	86.8	75.7	81.9	88.7	78.0	52.9	53.1	35.7	87.5	89.5	83.7	62.3	64.4	51.6
SPIICE (TIP22) [41]	81.7	90.8	81.2	73.4	83.8	70.5	44.8	46.8	29.4	82.8	92.1	83.6	57.2	64.6	47.9
SeCu (ICCV23) [48]	70.7	81.4	65.7	79.9	88.5	78.2	51.6	51.6	36.0	—	—	—	—	—	—
DivClust (CVPR23) [37]	—	—	—	71.0	81.5	67.5	44.0	43.7	28.3	85.0	90.0	81.9	51.6	52.9	37.6
RPSC (AAAI24) [35]	83.8	92.0	83.4	75.4	85.7	73.1	47.6	51.8	34.1	83.0	92.7	85.8	55.2	64.0	46.5
CLIP ( <i>k</i> -means)	91.7	94.3	89.1	70.3	74.2	61.6	49.9	45.5	28.3	96.9	98.2	96.1	39.8	38.1	20.1
SIC (AAAI23) [1]	95.3	98.1	95.9	<b>84.7</b>	<b>92.6</b>	<b>84.4</b>	59.3	58.3	43.9	97.0	98.2	96.1	69.0	69.7	55.8
TAC (ICML24) [32]	92.3	94.5	89.5	80.8	90.1	79.8	60.7	55.8	42.7	97.5	98.6	97.0	75.1	75.1	63.6
GradNorm (ours)	<b>95.6</b>	<b>98.3</b>	<b>96.2</b>	82.6	91.1	81.5	<b>61.3</b>	<b>60.6</b>	<b>43.6</b>	<b>98.7</b>	<b>99.4</b>	<b>98.7</b>	<b>81.0</b>	<b>81.2</b>	<b>70.9</b>

Table 2. Clustering performance (%) on three challenging image clustering datasets. The best results are highlighted in bold.

Dataset	DTD			UCF-101			ImageNet-1K			Average		
Metrics	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
CLIP (zero-shot)	56.5	43.1	26.9	79.9	63.4	50.2	81.0	63.6	45.4	72.5	56.7	40.8
SCAN (ECCV20) [56]	59.4	46.4	31.7	79.7	61.1	53.1	74.7	44.7	32.4	71.3	50.7	39.1
CLIP ( <i>k</i> -means)	57.3	42.6	27.4	79.5	58.2	47.6	72.3	38.9	27.1	69.7	46.6	34.0
SIC (AAAI23) [1]	59.6	45.9	30.5	81.0	61.9	53.6	77.2	47.0	34.3	72.6	51.6	39.5
TAC (ICML24) [32]	60.1	45.9	29.0	81.6	61.3	52.4	77.8	48.9	36.4	73.2	52.0	39.3
GradNorm (ours)	<b>63.1</b>	<b>50.9</b>	<b>34.2</b>	<b>82.9</b>	<b>62.7</b>	<b>53.2</b>	<b>79.2</b>	<b>52.6</b>	<b>39.1</b>	<b>74.9</b>	<b>55.4</b>	<b>41.7</b>

### 5.1.3. Evaluation Metrics

We measure clustering performance by three popular metrics, including Accuracy (ACC), Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI).

## 5.2. Main results

### 5.2.1. Performance on Classical Datasets

We evaluate our proposed GradNorm on five widely-used image clustering datasets, compared with 21 deep clustering baselines. Significantly different from early baselines adopt either ResNet-34 or ResNet-18 as the backbone, this paper mainly focuses on comparisons with zero-shot CLIP and CLIP-based methods. As shown in Table 1, GradNorm

consistently outperforms the mostly recent TAC [1] on 5 classic datasets. In particular, GradNorm achieves a notable 7.3% and 6.1% improvement in ARI and ACC on ImageNet-Dogs respectively, which eposes its theoretical superiority in Section 4. While SIC [32] slightly outperforms GradNorm on the CIFAR-10 dataset, it is worth mentioning that SIC [32] requires more trainable parameters and a more sophisticated training strategy.

### 5.2.2. Performance on Challenging Datasets

Considering that the rapid development of network pre-training has made clustering on relatively simple datasets such as STL-10 and CIFAR-10 longer challenging, we evaluate GradNorm on three challenging datasets with larger

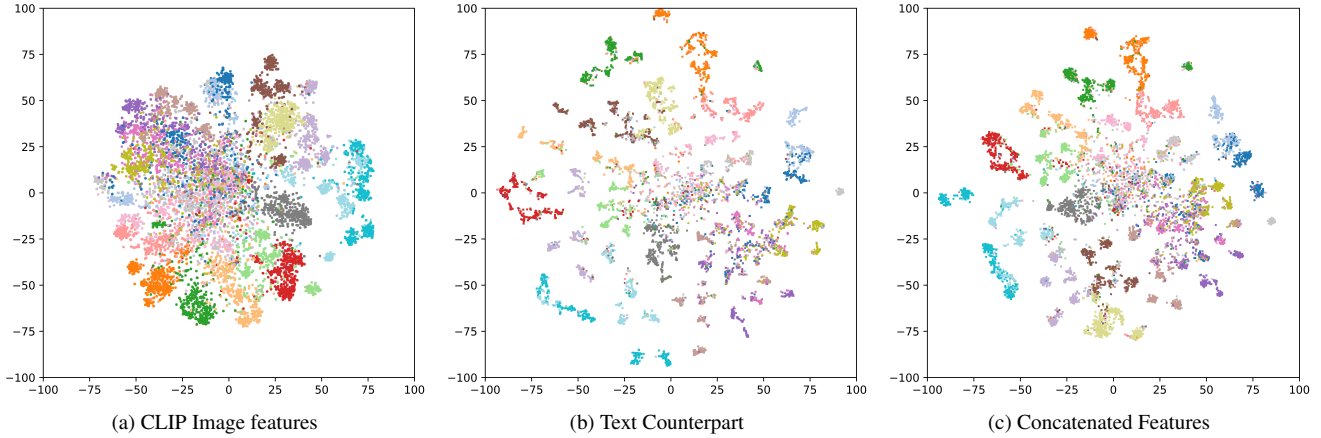


Figure 1. t-SNE Visualization of features extracted by different methods on the test split of CIFAR-20: (a) image embedding directly extracted from the pre-trained CLIP visual encoder, (b) text counterparts constructed by the candidate semantics selected by GradNorm, and (c) concatenation of images and text counterparts. Various colors indicate different ground-truth class assignment.

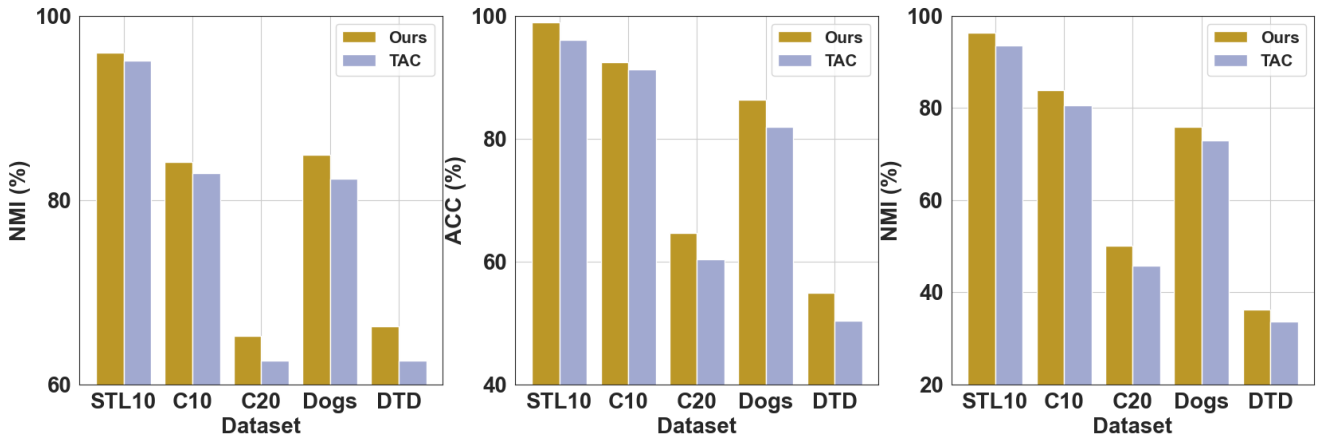


Figure 2. Clustering performance on five image clustering datasets, where CLIP-B/16 is used as encoder.

cluster numbers. Table 2 depicts the results on three challenge datasets, where our method still achieves the best performance. To be specific, our GradNorm outperforms TAC [32] over 5.0% ACC and 5.2% ARI on DTD. Besides, our method also outperforms supervised zero-shot CLIP, which highlights the effectiveness of our approach in applying CLIP for clustering tasks.

### 5.3. Visualizations

To provide an intuitive understanding of our empirical superiority in clustering, we present t-SNE [55] visualization on various features obtained by our GradNorm. Compared with the pre-trained CLIP image features in Figure 5a that suffers from remarkable overlapping among image features of different classes, the constructed text counterpart in Figure 5b exhibit better separation among clusters. Finally, Figure 5b implies that simply concatenating images

and text counterparts could better collaborate the image and text modalities, achieving the best trade-off between within-clustering compactness and between-cluster separation.

## 5.4. Ablation Study

### 5.4.1. Ablation on Hyper-parameters

We evaluate the hyper-parameters most essential to the algorithmic design of our GradNorm. To assess the impact of the temperature hyper-parameter  $\kappa$  in Eq. (9), we vary the value of  $\kappa$  from 0.002 to 0.02. The resulting clustering performance on UTD and CIFAR-20 is reported in Figure 4a and Figure 4b respectively. To assess the impact of the temperature hyper-parameter  $\tau$  in Eq. (4) and Eq. (6), we vary the value of  $\tau$  from 5 to 100. The resulting clustering performance on UTD and CIFAR-20 is reported in Figure 5a and Figure 5b respectively. As illustrated in Figure 6a and Figure 6b, the clustering performance of GradNorm exhibits

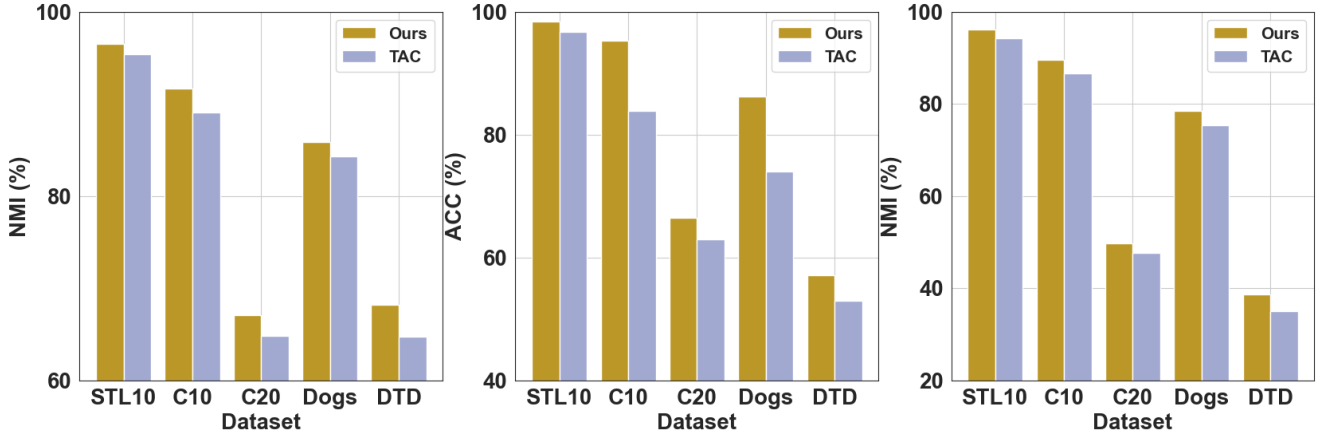


Figure 3. Clustering performance on five image clustering datasets, where CLIP-L/14 is used as encoder.

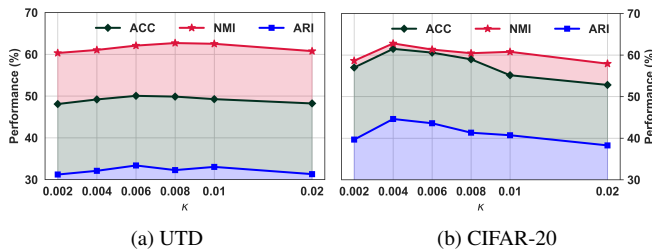


Figure 4. Analysis of clustering performance by varying the value of the temperature hyper-parameter  $\kappa$  on (a) UTD and (b) CIFAR-20 datasets, respectively.

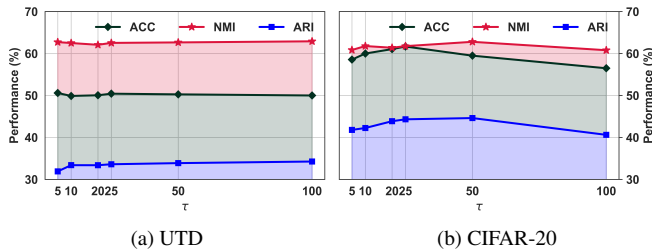


Figure 5. Analysis of clustering performance by varying the value of the temperature hyper-parameter  $\tau$  on (a) UTD and (b) CIFAR-20 datasets, respectively.

an initial improvement as  $\beta$  increases, followed by either reaching a stable level or degrading slightly when  $\beta$  is too high. We suspect that incorporating excessive nouns can introduce unrelated semantics, which has an adverse effect on the clustering process.

#### 5.4.2. Ablation on Visual Encoder

In principle, our GradNorm is generic to the choice of visual encoder. We evaluate GradNorm with different visual encoder architectures, including ViT-B/16 and ViT-L/14, and report the corresponding clustering results in Figure 2 and

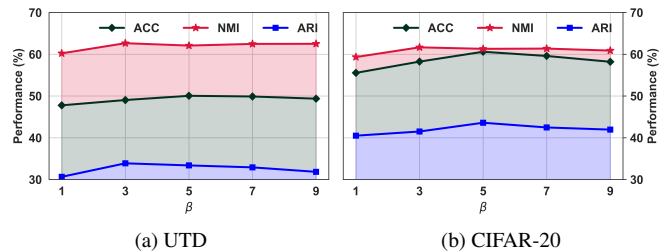


Figure 6. Analysis of clustering performance by varying  $\beta$ , the number of selected positive semantics, on (a) UTD and (b) CIFAR-20 datasets, respectively.

Figure 3. On the one hand, the clustering performance can be enhanced by more powerful visual encoders. On the other hand, GradNorm consistently outperforms TAC regardless of the backbone architecture used, which implies the better generalization of GradNorm over TAC.

## 6. Conclusion

In this paper, we propose a novel gradient-based framework GradNorm that exploits the unlabeled in-the-wild textual data for LaIC. Theoretically, GradNorm answers the question of how does unlabeled wild data help LaIC by analyzing the separability of truly positive semantics in the wild. Empirically, GradNorm achieves strong performance compared to competitive baselines on various datasets, which echoes our theoretical insights. Besides, extensive ablations provide further understandings of our GradNorm.

## Acknowledgement

This work is supported by the Australian Research Council Discovery Early Career Researcher Award (DE250100363) and the Australian Laureate Fellowship (FL190100149).

## References

- [1] Shaotian Cai, Liping Qiu, Xiaojun Chen, Qin Zhang, and Longteng Chen. Semantic-enhanced image clustering. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6869–6878, 2023. 1, 2, 3, 4, 5, 6
- [2] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018. 1
- [3] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *Proceedings of the IEEE international conference on computer vision*, pages 5879–5887, 2017. 1, 5, 6
- [4] Chunchun Chen, Wenjie Zhu, and Bo Peng. Differentiated graph regularized non-negative matrix factorization for semi-supervised community detection. *Physica A: Statistical Mechanics and its Applications*, 604:127692, 2022. 1
- [5] Chunchun Chen, Wenjie Zhu, Bo Peng, and Huijuan Lu. Towards robust community detection via extreme adversarial attacks. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2231–2237. IEEE, 2022. 1
- [6] Zixiang Chen, Yihe Deng, Yuanzhi Li, and Quanquan Gu. Understanding transferable representation learning and zero-shot transfer in clip. *arXiv preprint arXiv:2310.00927*, 2023. 3
- [7] Zixiang Chen, Yihe Deng, Yuanzhi Li, and Quanquan Gu. Understanding transferable representation learning and zero-shot transfer in clip. *arXiv preprint arXiv:2310.00927*, 2023. 1
- [8] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 5
- [9] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011. 5
- [10] Zhiyuan Dang, Cheng Deng, Xu Yang, Kun Wei, and Heng Huang. Nearest neighbor matching for deep clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13693–13702, 2021. 1, 6
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5
- [12] Kien Do, Truyen Tran, and Svetha Venkatesh. Clustering by maximizing mutual information across views. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9928–9938, 2021. 1, 6
- [13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5
- [14] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996. 2
- [15] Kamran Ghasedi Dizaji, Amirhossein Herandi, Cheng Deng, Weidong Cai, and Heng Huang. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In *Proceedings of the IEEE international conference on computer vision*, pages 5736–5745, 2017. 2
- [16] Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *International conference on machine learning*, pages 1558–1567. PMLR, 2017. 2
- [17] Jiabo Huang, Shaogang Gong, and Xi Tian Zhu. Deep semantic clustering by partition confidence maximisation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8849–8858, 2020. 1, 6
- [18] Zhizhong Huang, Jie Chen, Junping Zhang, and Hongming Shan. Learning representation for clustering via prototype scattering and positive sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7509–7524, 2022. 1, 2
- [19] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992. 3
- [20] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9865–9874, 2019. 2, 6
- [21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2
- [22] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. *arXiv preprint arXiv:1611.05148*, 2016. 1, 2
- [23] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021. 2
- [24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009. 5
- [26] Marc T Law, Raquel Urtasun, and Richard S Zemel. Deep spectral clustering learning. In *International conference on machine learning*, pages 1985–1994. PMLR, 2017. 1
- [27] Ted Lentsch, Holger Caesar, and Dariu Gavrilă. Union: Unsupervised 3d object detection using object appearance-based pseudo-classes. *Advances in Neural Information Processing Systems*, 37:22028–22046, 2025. 1

- [28] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2
- [29] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8547–8555, 2021. 1, 2, 6
- [30] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 2
- [31] Yunfan Li, Mouxing Yang, Dezhong Peng, Taihao Li, Jiantao Huang, and Xi Peng. Twin contrastive learning for online clustering. *International Journal of Computer Vision*, 130(9):2205–2221, 2022. 1, 2, 6
- [32] Yunfan Li, Peng Hu, Dezhong Peng, Jiancheng Lv, Jianping Fan, and Xi Peng. Image clustering with external guidance. *arXiv preprint arXiv:2310.11989*, 2023. 1, 2, 3, 4, 5, 6, 7
- [33] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):171–184, 2012. 2
- [34] Honglin Liu, Peng Hu, Changqing Zhang, Yunfan Li, and Xi Peng. Interactive deep clustering via value mining. *Advances in Neural Information Processing Systems*, 37:42369–42387, 2024. 1
- [35] Sihang Liu, Wenming Cao, Ruigang Fu, Kaixiang Yang, and Zhiwen Yu. Rpsc: robust pseudo-labeling for semantic clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14008–14016, 2024. 6
- [36] Yiding Lu, Haobin Li, Yunfan Li, Yijie Lin, and Xi Peng. A survey on deep clustering: from the prior perspective. *Vicinityearth*, 1(1):4, 2024. 1
- [37] Ioannis Maniatis Metaxas, Georgios Tzimiropoulos, and Ioannis Patras. Divclust: Controlling diversity in deep clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3418–3428, 2023. 6
- [38] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 1, 3, 5
- [39] Sudipto Mukherjee, Himanshu Asnani, Eugene Lin, and Sreeram Kannan. ClusterGAN: Latent space clustering in generative adversarial networks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4610–4617, 2019. 1, 2
- [40] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015. 5
- [41] Chuang Niu, Hongming Shan, and Ge Wang. Spice: Semantic pseudo-labeling for image clustering. *IEEE Transactions on Image Processing*, 31:7264–7278, 2022. 6
- [42] Bo Peng and Wenjie Zhu. Deep structural contrastive subspace clustering. In *Asian Conference on Machine Learning*, pages 1145–1160. PMLR, 2021. 2
- [43] Bo Peng, Wenjie Zhu, and Xiuhui Wang. Deep residual matrix factorization for gait recognition. In *Proceedings of the 2020 12th International Conference on Machine Learning and Computing*, pages 330–334, 2020. 1
- [44] Bo Peng, Zhen Fang, Guangquan Zhang, and Jie Lu. Knowledge distillation with auxiliary variable. In *Forty-first International Conference on Machine Learning*, 2024.
- [45] Bo Peng, Yadan Luo, Yonggang Zhang, Yixuan Li, and Zhen Fang. Conjnorm: Tractable density estimation for out-of-distribution detection. *arXiv preprint arXiv:2402.17888*, 2024.
- [46] Bo Peng, Jie Lu, Yonggang Zhang, Guangquan Zhang, and Zhen Fang. Distributional prototype learning for out-of-distribution detection. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 1104–1114, 2025. 1
- [47] Xi Peng, Shijie Xiao, Jiashi Feng, Wei-Yun Yau, and Zhang Yi. Deep subspace clustering with sparsity prior. In *Ijcai*, pages 1925–1931, 2016. 2
- [48] Qi Qian. Stable cluster discrimination for deep clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16645–16654, 2023. 4, 5, 6
- [49] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1, 2
- [50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 5
- [51] Yuming Shen, Ziyi Shen, Menghan Wang, Jie Qin, Philip Torr, and Ling Shao. You never cluster alone. *Advances in Neural Information Processing Systems*, 34:27734–27746, 2021. 4, 6
- [52] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5
- [53] Yaling Tao, Kentaro Takagi, and Kouta Nakata. Clustering-friendly representation learning via instance discrimination and feature decorrelation. *arXiv preprint arXiv:2106.00131*, 2021. 2, 6
- [54] Tsung Wei Tsai, Chongxuan Li, and Jun Zhu. Mice: Mixture of contrastive experts for unsupervised image clustering. In *International conference on learning representations*, 2020. 2, 5, 6
- [55] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7
- [56] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European conference on computer vision*, pages 268–285. Springer, 2020. 1, 6
- [57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia

- Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5
- [58] Zhen Wang, Zhaoqing Li, Rong Wang, Feiping Nie, and Xuelong Li. Large graph clustering with simultaneous spectral embedding and discretization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4426–4440, 2020. 2
- [59] Jianlong Wu, Keyu Long, Fei Wang, Chen Qian, Cheng Li, Zhouchen Lin, and Hongbin Zha. Deep comprehensive correlation mining for image clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8150–8159, 2019. 6
- [60] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 5
- [61] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016. 1, 2, 6
- [62] Bo Yang, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *international conference on machine learning*, pages 3861–3870. PMLR, 2017. 1
- [63] Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5147–5156, 2016. 6
- [64] Xu Yang, Cheng Deng, Feng Zheng, Junchi Yan, and Wei Liu. Deep spectral clustering using dual autoencoder network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4066–4075, 2019. 1
- [65] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 2
- [66] Chunlin Yu, Ye Shi, and Jingya Wang. Contextually affine neighborhood refinery for deep clustering. *Advances in Neural Information Processing Systems*, 36:5778–5790, 2023. 1
- [67] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017. 4
- [68] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 2
- [69] Yonggang Zhang, Jie Lu, Bo Peng, Zhen Fang, and Yiu-ming Cheung. Learning to shape in-distribution feature space for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 37:49384–49402, 2024. 1
- [70] Huasong Zhong, Jianlong Wu, Chong Chen, Jianqiang Huang, Minghua Deng, Liqiang Nie, Zhouchen Lin, and Xian-Sheng Hua. Graph contrastive clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9224–9233, 2021. 1, 6
- [71] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 2
- [72] Qinli Zhou, Wenjie Zhu, Hao Chen, and Bo Peng. Community detection in multiplex networks by deep structure-preserving non-negative matrix factorization. *Applied Intelligence*, 55(1):26, 2025. 1
- [73] Wenjie Zhu and Bo Peng. Sparse and low-rank regularized deep subspace clustering. *Knowledge-Based Systems*, 204: 106199, 2020. 2
- [74] Wenjie Zhu and Bo Peng. Manifold-based aggregation clustering for unsupervised vehicle re-identification. *Knowledge-Based Systems*, 235:107624, 2022. 1
- [75] Wenjie Zhu, Bo Peng, Han Wu, and Binhao Wang. Query set centered sparse projection learning for set based image classification. *Applied Intelligence*, 50(10):3400–3411, 2020. 1
- [76] Wenjie Zhu, Bo Peng, and Chunchun Chen. Self-supervised embedding for subspace clustering. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3687–3691, 2021. 2
- [77] Wenjie Zhu, Chunchun Chen, and Bo Peng. Unified robust network embedding framework for community detection via extreme adversarial attacks. *Information Sciences*, 643:119200, 2023. 1
- [78] Wenjie Zhu, Bo Peng, Chunchun Chen, and Hao Chen. Deep discriminative dictionary pair learning for image classification. *Applied Intelligence*, 53(19):22017–22030, 2023.
- [79] Wenjie Zhu, Bo Peng, and Wei Qi Yan. Dual knowledge distillation on multiview pseudo labels for unsupervised person re-identification. *IEEE Transactions on Multimedia*, 26: 7359–7371, 2024. 1