Deep Clustering by Semantic Contrastive Learning

Jiabo Huang
jiabo.huang@qmul.ac.uk
Shaogang Gong
s.gong@qmul.ac.uk

Computer Vision Group,
School of Electronic Engineering and Computer Science,
Queen Mary University of London,
London, E1 4NS, UK

Abstract

Whilst contrastive learning has recently brought notable benefits to deep clustering of unlabelled images by learning sample-specific discriminative visual features, its potential for explicitly inferring class decision boundaries is less well understood. This is because its instance discrimination strategy is not class sensitive, hence, the clusters derived on the resulting feature space are not optimised for corresponding to meaningful class decision boundaries. In this work, we solve this problem by introducing Semantic Contrastive Learning (SCL). SCL imposes explicitly distance-based cluster structures on unlabelled training data by formulating a semantic (cluster-aware) contrastive learning objective. Specifically, we encourage consensus between learning the optimal hypotheses on the semantic class boundaries and feature similarities. This is formulated by a clustering consistency condition to be satisfied jointly by *instance* feature similarities and *cluster* decision boundaries. This semantic contrastive learning approach to discovering unknown class decision boundaries has considerable advantages to unsupervised learning of object recognition. Extensive experiments show that SCL outperforms state-of-the-art contrastive learning and deep clustering methods on six object recognition benchmarks, especially on the more challenging finer-grained and larger datasets.

1 Introduction

Given the massive increase of images available on the Internet, how to leverage them without label annotation for learning high-level visual semantics remains a challenging problem for unsupervised deep learning, although it has been shown to be highly effective in supervised deep learning given large-scale labelled training data. Clustering as a conventional unsupervised machine learning technique [1, 21, 28] has been recently exploited for visual representation learning in deep neural networks to perform *Deep Clustering* [15, 20].

Separately, contrastive learning [8, 14, 36] has also been shown effective for self-supervised learning of *generalisable* feature representations by *instance-discrimination* (Fig. 1 (a)). It may appear to have the potential to benefit unsupervised clustering due to its expressive sample-specific visual representation. However, directly applying contrastive learning to deep clustering is sub-optimal for *class-discrimination*. Because it lacks awareness of non-linear intra-class variations. This is inherent from learning with per-sample pseudo classes

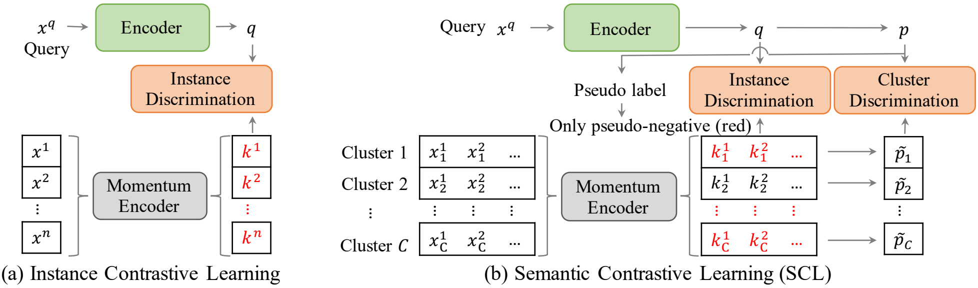


Figure 1: Instance *vs*. semantic contrastive learning. (a) Instance contrastive learning differentiates query samples from a contrastive set regardless their potential class memberships. (b) SCL pulls samples away from only their pseudo negatives of other clusters.

generated by global linear data augmentations for instance-wise visual discrimination. We observe that this limitation is overlooked by recent contrastive learning based clustering methods. Such attempts avoid the problem by either limiting to restricted local feature space neighbourhoods exhibiting subtle visual variations [16, 17], or suffering from class ambiguities due to the inherent contradiction between instance-level discrimination (pull away intra-cluster) and class(cluster)-level grouping (push closer intra-cluster) [27, 35, 37].

In this work, we propose a deep clustering method called *Semantic Contrastive Learning* (SCL). In this SCL model, cluster structures are explicitly imposed to unlabelled training data to encourage learning a 'cluster-aware' instance discriminative feature space that promotes separation of decision boundaries between clusters, leading to a plausible interpretation of the underlying semantic concepts (Fig. 1 (b)). Specifically, the instance discrimination in SCL aims to reduce visual redundancy (what's common) *across samples* so that images which are sharing more uncommon (what's unique) appearance patterns are pushed closer in feature space, whilst the cluster discrimination aims to optimise holistically the cluster decision boundaries so that any visual overlap *across clusters* is minimised and each cluster exhibits unique and consistent visual characteristics of each underlying class (semantic concept). Different from the recent instance contrastive learning based clustering methods [27, 35, 37] which pull away each instance from *all* other samples in the feature space, SCL only pulls it away from its *pseudo-negative* samples in other clusters. By sharing a common contrastive (negative) set for all the instances in a cluster, SCL indirectly pushes them closer regardless of any intra-cluster visual dissimilarity. This resolves the contradiction in the instance contrastive learning and clustering objectives but is neglected by those recent attempts. Moreover, we introduce a new semantic memory to not only store representations for instance discrimination but also embed the cluster structures. This enables optimising cluster decision boundaries by maximising the consistency between cluster-level (semantic) and instance-level (visual) distances.

Our **contributions** are: **(1)** We make the first attempt to solve the contradiction in learning simultaneously instance contrastive discrimination and clustering objectives in order to optimise nontrivial class separations in a feature space without labelled training. **(2)** We introduce a novel *Semantic Contrastive Learning* (SCL) for deep clustering. SCL discovers cluster decision boundaries by enforcing a consensus between instance contrastive discrimination and cluster compactness. **(3)** We formulate a new semantic memory to enable simultaneous optimisation of instance and cluster discrimination. SCL yields compelling performance advantages over the state-of-the-art deep clustering methods, with significant improvements (∼17%) on the more challenging larger and finer-grained datasets.

2 Related Work

We shall first differentiate clearly the different objectives between deep clustering in the context of this work and unsupervised representation learning elsewhere. The latter aims to learn generalisable feature representations – generative representational learning – without any consideration for optimising class-discrimination. Our objective is generative decision boundary learning optimised for class-discrimination without labels in model learning.

Deep clustering. In the absence of ground-truth class labels, one popular solution of deep clustering is to mimic supervised learning by estimating pseudo labels iteratively from learning improvement on feature representations [6, 7, 12, 38, 40, 41, 42]. Although these methods may benefit from explicit supervised discriminative learning, it is also intrinsically unstable due to error-propagation between unreliable label assignments and updates of randomly initialised representation based on such assignments [16, 46]. SCL is more robust to error propagation from the intermediate cluster assignments during model learning because the contrastive learning formulation is able to discover the intrinsic visual similarity among samples despite a lack of knowledge of their true class memberships. In contrast to the alternate strategy, one can learn simultaneously label assignment and feature updates using certain pretext objectives that indirectly impose requirements for learning good cluster structures [13, 18, 19, 20, 31, 33, 44]. However, due to the weak correlations between their learning objectives and the target class boundary separations, they tend to yield clusters that are less consistent with the semantic categories. SCL reduces visual redundancy across clusters so that each cluster exhibits unique and consistent visual characteristics that are more plausible for encoding an underlying semantic concept.

There are a few recent attempts [11, 27, 35, 37] on deep clustering by exploring directly visual features from instance contrastive learning. However, they either suffer from class ambiguities due to the inherent contradiction between instance-level discrimination (pull away intra-cluster) and cluster-level grouping (push closer intra-cluster) [11, 27], or focused only on a one-sided representation learning [35] or their partitioning [37] while neglecting their mutual impacts. By assembling instance-wise contrastive samples into a common pseudo-negative set for simultaneous instance discrimination and cluster decision boundary optimisation, we resolve their contradiction and jointly amplify their strengths.

Unsupervised representation learning. Beyond the forementioned works designed for modelling the inherent class structure of unlabelled images, there are other methods for learning generalisable image representations that may appear to be similar to clustering [2, 3, 4, 5, 26]. Those representation learning methods assume clustering *is given*, which rely on independent clustering [3, 4, 26] or optimal transport algorithms [2, 5] to compute the pseudo labels. Therefore, they are both limited by potentially suboptimal clustering computed independently, and only addressing restricted partial problem, an easier learning task. Our SCL model solves the two underlying problems holistically as a single problem by focusing directly on modelling semantically-aware clustering therefore removing any suboptimal offline clustering assumption and is end-to-end optimised for the resulting representation derived.

Contrastive learning [8, 9, 14, 32, 36, 39] optimises sample-specific visual features by treating every individual instance as an independent class augmented by guaranteed positive samples generated using global linear transforms. By ignoring any cross-sample relationships and global class memberships, the learned representations are ambiguous to both intra and inter-class nonlinear image variations, therefore, less discriminative against true classes. To address such a limitation, studies have been carried out to integrate it with neighbourhood discovery [16, 17, 45]. These methods adopt directly the supervised contrastive learing [22] paradigm to *explicitly* push pseudo-positive samples *closer* in the feature space. Such a paradigm is prone to accumulating errors from unreliable pseudo label predictions. Extra constraints and strategies such as restricting neighbourhood's size and pre-learning representations must be applied to avoid the negative impacts of error-propagation. Such strategies cannot apply in general therefore are suboptimal. In contrast, our SCL model *implicitly* poses positive relationships by pulling samples *away* from a common pseudo-negative contrastive set. SCL has no need for hand-crafted extra strategies which are time consuming and non-scalable due to being independent from the deep clustering learning, not end-to-end.

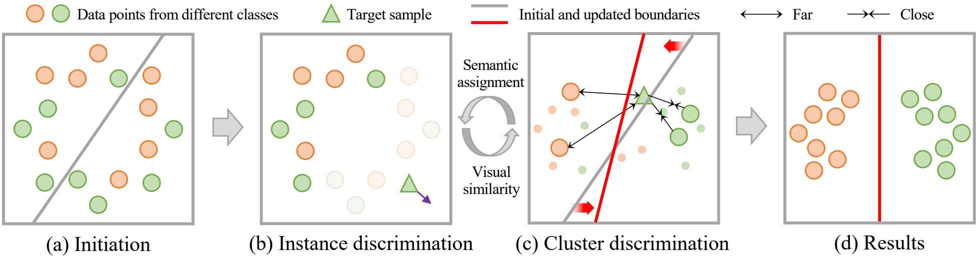

Data points from different classes
Target sample
Initial and updated boundaries
Far
Close
Semantic assignment
Visual similarity
(a) Initiation
(b) Instance discrimination
(c) Cluster discrimination
(d) Results

Figure 2: An overview of SCL. **(a)** Given a randomly initialised feature space and decision boundaries, **(b)** the SCL model optimises visual similarities among samples by instance discrimination and **(c)** the potential class memberships by cluster discrimination **(d)** and finally converge to a consensus between instance-level diversity and cluster-level compactness.

3 Clustering on Unlabelled Images

Given a set of *unlabelled* images $\mathcal{I} = \{\boldsymbol{I}_1, \boldsymbol{I}_2, \cdots, \boldsymbol{I}_N\}$, deep clustering aims to derive (1) a *feature embedding network* $\boldsymbol{\theta}$ that extracts key semantic information encoded in the high-dimensional pixel space to a compact vector subspace $f_\theta : \boldsymbol{I} \rightarrow \boldsymbol{x} \in \mathbb{R}^d$, and (2) a *classifier* ϕ that projects the feature vectors into C partitions $f_\phi : \boldsymbol{x} \rightarrow y, y \in \{1, 2, \cdots C\}$, with a hope that samples in the same cluster share the same ground-truth class label, otherwise not. It is fundamentally challenging to derive class discriminative information directly from raw images in an unsupervised manner, due to the complex appearance patterns and variations exhibited both within and across classes.

In this work, we introduce a *Semantic Contrastive Learning* (SCL) method. SCL explores the idea of cluster-centred contrastive learning that differ from other recent developments on deep clustering by directly applying instance-centred contrastive learning. Given the sample-specific learning constraint of the instance contrastive learning, it is non-trivial to exploit it in unsupervised clustering that also jointly enforces necessary constraints to unknown class decision boundary when there is no class label in training. To overcome this hurdle, we optimises concurrently instance discrimination and their assignment to a set of clusters, with an additional consistency objective function to condition their optimisations jointly. SCL aims to learn both optimal instance visual similarities that can verify each instance's cluster assignment globally and optimal cluster compactness that can maximise inter-cluster discrimination margins. Importantly, the SCL formulation can be utilised by any instance contrastive learning methods [8, 14, 39] for deep clustering tasks, and it is end-to-end trainable therefore globally optimised. Fig. 2 shows an overview of SCL.

3.1 Semantic Contrastive Learning

We start with formulating a new *cross-cluster* instance discrimination learning objective with a novel *semantic memory*. The aim is to learn visual features to be discriminative across clusters and facilitate simultaneous instance and cluster discrimination.

Cross-cluster instance discrimination. Our feature learning objective is formulated to differentiate every individual instance against its pseudo-negative samples so to reduce its visual redundancy regarding images of other clusters (Fig. 2 (b)). Given random partitions at the beginning of training (Fig. 2 (a)), by isolating samples from different clusters, the model behaves as instance contrastive learning and outputs per sample-specific visual features. Intuitively, visually similar samples are expected to share more class-specific unique information, their representations will therefore be gradually gathered closer and grouped into the same clusters by our cluster discrimination detailed later. Along the clustering process with increasingly better and stable cluster assignments, the contrastive set of every sample will absorb more visually dissimilar counterparts, instead of random ones. Consequently, the learning objective becomes reducing cross-cluster visual redundancy, resulting in desired features that are aware of inter-cluster visual discrepancies and invariant within clusters (Fig. 2 (d)).

Whilst our SCL is a generic formulation, we take the momentum contrast (MoCo) [9, 14] as an example of instantiation. We first formulate a mapping function f_θ from a pixel space to a representational space as an encoder with learnable weights θ. Similarly, we construct another momentum encoder $f_{\tilde{\theta}}$ with an identical structure but independent parameters $\tilde{\theta}$. Given an unlabelled dataset $\mathcal{I}$, we randomly apply a set of transformations $\mathcal{T}$ to each image for distribution perturbation. We then represent two perturbed copies of each instance, $\mathcal{T}_1(\boldsymbol{I}_i)$ and $\mathcal{T}_2(\boldsymbol{I}_i)$, by the two encoders respectively and denote them as $\boldsymbol{q}_i = f_\theta(\mathcal{T}_1(\boldsymbol{I}_i))$ and $\boldsymbol{k}_i = f_{\tilde{\theta}}(\mathcal{T}_2(\boldsymbol{I}_i))$. Given the pseudo labels of all the samples $\mathcal{Y} = \{y_1, y_2, \cdots, y_N\}$, $y_i \in [1, C]$ inferred by the progressively updating decision boundaries, our instance discrimination objective in terms of $\boldsymbol{I}_i$ is to match $\boldsymbol{q}_i$ with $\boldsymbol{k}_i$ against its contrastive set $Q_i = \{\tilde{\boldsymbol{k}}_1, \tilde{\boldsymbol{k}}_2, \cdots, \tilde{\boldsymbol{k}}_K\}\ s.t.\ y_i \neq y_j, \forall j \in [1, K]$ composed by K stale representations of its pseudo-negative samples:

$$\mathcal{L}_{\text{ID}}(\boldsymbol{I}_i) = -\log \frac{\exp(cos(\boldsymbol{q}_i, \boldsymbol{k}_i)/\tau)}{\sum_{\tilde{\boldsymbol{k}} \in Q_i \cup \{\boldsymbol{k}_i\}} \exp(cos(\boldsymbol{q}_i, \tilde{\boldsymbol{k}})/\tau)}, \tag{1}$$

where $cos(\cdot, \cdot)$ is the cosine similarity between a pair of representations and τ is the temperature to control the concentration degree of distribution. As the samples in the same clusters share a common contrastive set, they are indirectly pushed closer in the feature space regardless of any intra-cluster variations. Therefore, the learned features are geared towards being sensitive to cluster-wise visual characteristics, not sample-wise.

Semantic memory. To facilitate instance discrimination across clusters, we manage C independent memory banks $\mathcal{M} = \{M_1, M_2, \cdots M_C\}$ each corresponding to one cluster with a size of $K/(C-1)$. For an image $\boldsymbol{I}_i$ with pseudo label y_i, we construct its contrastive set Q_i:

$$Q_i = \{\tilde{\boldsymbol{k}} | \tilde{\boldsymbol{k}} \in M_j\ \forall j \in [1, C] \text{ and } j \neq y_i\}. \tag{2}$$

There is always one memory bank left out for each sample and the rest Ms are concatenated as its contrastive set Q approximately in size K (rounding error) to support cluster discriminative feature representation learning. For memory update, after every backward pass, the representation $\boldsymbol{k}_i$ enqueues to M_{y_i} with the oldest one inside removed.

Cluster discrimination. To discover the underlying concepts with unique visual characteristics, we infer their decision boundaries by reducing the visual redundancy among clusters, namely maximising the visual similarity of samples within the same clusters and minimising that between clusters (Fig. 2 (c)). Concretely, as the representation of samples with different pseudo labels are stored independently in the semantic memory bank, they can be taken as anchors to describe their corresponding clusters. Given a training sample $\boldsymbol{q}_i$, its probability $\tilde{p}_{i,j}$ of being in the j-the cluster predicted by a distance-based classifier is

$$\tilde{p}_{i,j} = \frac{\sum_{\tilde{\boldsymbol{k}} \in M_j} \exp(cos(\boldsymbol{q}_i, \tilde{\boldsymbol{k}})/\tau)}{\sum_{j'=1}^{C} \sum_{\tilde{\boldsymbol{k}} \in M_{j'}} \exp(cos(\boldsymbol{q}_i, \tilde{\boldsymbol{k}})/\tau)}. \tag{3}$$

With such potential memberships determined by sample-anchor visual similarities, we formulate a consistency loss for learning the cluster decision boundaries:

$$\boldsymbol{p}_i = \text{Softmax}(W^\top \boldsymbol{q}_i + B) \in \mathcal{R}^C, \quad \mathcal{L}_{\text{CD}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{C} -\tilde{p}_{i,j} \log p_{i,j}, \tag{4}$$

where $\{W; B\}$ is the learnable parameters of classifier f_ϕ and n denotes the size of mini-batch. In Eq. (4), we aim to minimise the cross-entropy of the distance-based cluster assignments $\tilde{\boldsymbol{p}}_i$ and the predictions $\boldsymbol{p}_i$ yielded by the cluster decision boundaries then propagate the gradient back to $\boldsymbol{p}_i$ only to avoid feature learning from unreliable boundaries. By doing so, samples are assigned to the cluster with the most similar anchors while each cluster holding its own visual characteristics that make it different from others and correspond to an underlying semantic class with consistent and unique visual characteristics.

With the updated models f_θ and f_ϕ, we renew the cluster assignments every epoch in a maximum likelihood manner for semantic memory construction in Eq. (2):

$$y_i = \arg\max_j \, p_{i,j}, \; j \in \{1, 2, \cdots, C\}. \tag{5}$$

As the predictions become increasingly more accurate in the process of training, this update improves cross-cluster instance discrimination on learning class discriminative features.

Hard samples mining. To enhance discrimination capacity, we identify semantically ambiguous samples and emphasise them in instance discrimination:

$$s_i^e = s_i^{e-1} + \mathbb{1}[y_i^e \neq y_i^{e-1}], \quad w_i^e = \frac{s_i^e}{\sum_j^n s_j^e}, \quad \mathcal{L}_{\text{ID}} = \sum_{i=1}^{n} w_i^e \mathcal{L}_{\text{ID}}(\boldsymbol{I}_i), \tag{6}$$

where w_i^e is the weights of $\boldsymbol{I}_i$ at the e-th training epoch. The samples that are frequently swapped across clusters (*i.e.* hard samples) are assigned with higher weights for offering more useful discriminative learning clues.

3.2 Model Training

Given the instance (Eq. (6)) and cluster (Eq. (4)) discrimination losses, the overall training objective of SCL is:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{ID}} + \beta \mathcal{L}_{\text{CD}}. \tag{7}$$

In the absence of labelled validation data in unsupervised clustering, we set both the weights to $\alpha = \beta = 1$ to avoid exhaustive per-dataset parameter tunning. To minimise $\mathcal{L}$, the weights of encoder θ as well as the decision boundaries ϕ are updated by back-propagation and the momentum encoder $\tilde{\theta}$ is by $\tilde{\theta} \leftarrow m\tilde{\theta} + (1-m)\theta$ where m is a momentum coefficient [14]. Both objective functions (Eq. (6) and Eq. (4)) are differentiable thus can be trained end-to-end by the conventional stochastic gradient descent algorithm.

4 Experiments

Datasets. Evaluations were conducted on six challenging object recognition benchmarks. **(1) CIFAR-10(/100) [23]**: Natural image datasets composed by 60,000 samples that are uniformly drawn from 10(/100) classes. The 20 super-classes on CIFAR-100 were considered as ground-truth. **(2) STL-10 [10]**: An ImageNet adapted dataset consists of 1,300 images from each of 10 classes. Additional 100,000 images from unknown classes were available but deprecated in our experiments. **(3) ImageNet-10/Dogs [34]**: ImageNet subsets containing samples from 10 randomly selected classes or 15 dog breeds. **(4) Tiny-ImageNet [25]**: Another ImageNet subset in larger-scale with 100,000 samples evenly distributed in 200 classes. Training and testing are conducted on the same set of data following convention [19, 20].

Evaluation metrics. Three standard clustering metrics were used to measure the consistency of cluster assignments and ground-truth class memberships: (1) Clustering accuracy (**ACC**) maps one-to-one the learned clusters to the ground-truth classes by the Hungarian algorithm [24] and measures the classification accuracy; (2) Normalised mutual information (**NMI**) quantifies the labelling consistency by the normalised MI between the predicted and ground-truth labels of all image samples; (3) Adjusted rand index (**ARI**) computes the ratio of image sample pairs holding consistent pairwise relationships against the ground-truth. All these metrics scale from 0 to 1 and higher is better.

Implementation details. We followed [19, 20] to use a variant of ResNet-34 as the backbone network and [9] for the other implementation choices. All our models and the cluster assignments are randomly initialised. An SGD optimiser was adopted for model updates with weight decay in $5e-4$. The coefficient for momentum encoder updating was 0.9 and τ in Eq. (1) was 0.1. We stored $4096/(C-1)$ representations for each cluster in the semantic memory (Eq. (2)) on all the datasets except for $8192/(C-1)$ on Tiny-ImageNet due to larger scale. The learning rate was set to 0.03 with the cosine schedule [29] for its adjustment across 200 epochs while the batch size was 256. We adopted the “merge-and-split” strategy [43] for updating pseudo labels (Eq. (5)) to avoid extremely imbalanced partitions and to stabilise training. Besides the target ‘clustering’ tasks which partition the target data into the ground-truth number of clusters to facilitate comparisons, we followed [19, 20] to jointly train SCL with auxiliary ‘under-clustering’ and ‘over-clustering’ tasks so to explore multi-grained visual similarity. The cluster number in ‘under-clustering’ was half of the ground-truth while instance-wise learning was considered as extreme ‘over-clustering’. At test time, we followed [27, 35] to compare by the best models using the assignments yielded by the classifier for ‘clustering’ tasks while the other two were deprecated. All the hyper-parameters were kept the same across different datasets, *i.e.* no exhaustive per dataset tuning. On computational cost, the only extra parameters we introduced to MoCo [9] are in the linear classifier f_ϕ and it took around 30 seconds on CIFAR-10 to update pseudo labels per epoch.

Dataset	CIFAR-10			CIFAR-100			STL-10			ImageNet-10			ImageNet-Dogs			Tiny-ImageNet		
Metrics	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
K-means	.087	.229	.049	.084	.130	.028	.125	.192	.061	.119	.241	.057	.055	.105	.020	.065	.025	.005
DEC* [40]	.257	.301	.161	.136	.185	.050	.276	.359	.186	.282	.381	.203	.122	.195	.079	.115	.037	.007
DAC* [6]	.396	.522	.306	.185	.238	.088	.366	.470	.257	.394	.527	.302	.219	.275	.111	.190	.066	.017
ADC* [13]	-	.325	-	-	.160	-	-	.530	-	-	-	-	-	-	-	-	-	-
DDC* [7]	.424	.524	.329	-	-	-	.371	.489	.267	.433	.577	.345	-	-	-	-	-	-
DCCM* [38]	.496	.623	.408	.285	.327	.173	.376	.482	.262	.608	.710	.555	.321	.383	.182	.224	.108	.038
IIC [19]	.513	.617	.411	-	.257	-	.431	.499	.295	-	-	-	-	-	-	-	-	-
PICA [20]	.591	.696	.512	.310	.337	.171	.611	.713	.531	.802	.870	.761	.352	.352	.201	.277	.098	.040
DCCS* [44]	.569	.656	.469	-	-	-	.376	.482	.262	.608	.710	.555	-	-	-	-	-	-
GAT* [31]	.475	.610	.402	.215	.281	.116	.446	.583	.363	.594	.739	.552	.281	.322	.163	-	-	-
SCAN† [37]	.712	.818	.665	.441	.422	.267	.654	.755	.590	-	-	-	-	-	-	-	-	-
IDFD† [35]	.711	.815	.663	.426	.425	.264	.643	.756	.575	**.898**	**.954**	**.901**	.546	.591	.413	-	-	-
CC† [27]	.705	.790	.637	.431	.429	.266	**.764**	**.850**	**.726**	.859	.893	.822	.445	.429	.274	.340	.140	.071
CRLC† [11]	.679	.799	.634	.416	.425	.263	.729	.818	.682	.831	.854	.759	.461	.484	.297	-	-	-
GCC† [11]	**.764**	**.856**	**.728**	.472	.472	.305	.684	.788	.631	.842	.901	.822	.490	.526	.362	.347	.138	.075
SCL†*	.744	.813	.683	**.477**	**.482**	**.314**	.593	.638	.485	.877	.930	.861	**.728**	**.763**	**.652**	**.337**	**.172**	**.080**

Table 1: Comparisons to the state-of-the-art deep clustering approaches. Methods with $(\cdot)^\dagger$ conducted deep clustering by contrastive learning and $(\cdot)^*$ trained without the additional data on STL-10. The $1^{st}/2^{nd}$ best results are highlighted in **red**/blue.

4.1 Comparisons to the State-of-the-Art

Deep Clustering. Table 1 compares the proposed SCL with a wide range of state-of-the-art deep clustering models including both with- (from "SCAN" and below) and without- (from "GAT"and above) contrastive learning in their formulation. We observe: **(1)** SCL has broad advantages over all other methods including the few close competitors, *e.g.* by 17.2% (ACC) improvement over IDFD on ImageNet-Dogs. SCL yielded the best results in 4 out of the 6 benchmarks and at least top-2 in 5/6 benchmarks. On STL-10 where SCL seems less competitive, the top models used almost 10 times more additional training images that are sampled from the same distribution as the target data but are explicitly of different classes independent from the target classes. Those additional data give significant benefits by learning from strong negative signals but then were explicitly excluded when training the target classifier, making it an easier learning task. In our experiment, we avoided using such a data engineering strategy because it is neither practical nor scalable to have such similar and guaranteed negative data unless their class labels are available. On the other hand, SCL's performance advantages over those methods learned without the additional data engineering (marked with $*$) remain notable, *e.g.* improving GAT by 5.5%. This is a more accurate reflection on models' true performances which is also consistent to the other benchmarks. **(2)** It is always more challenging to precisely model the truth class boundaries of either finer-grained or larger datasets. In these cases, SCL surpassed IDFD and CC on ImageNet-Dogs and Tiny-ImageNet by 17.2% and 3.2%, respectively. **(3)** The significant performance margins obtained by all contrastive learning based methods indicate compellingly the benefit of contrastive constraints in unsupervised semantic concepts learning. Importantly, SCL's superiority demonstrate the significance of solving the contradiction between optimising instance contrastive discrimination (pull apart) and intra-cluster compactness (push closer).

Representation learning. Beyond the methods intrinsically designed for clustering [11, 27, 35, 37], we also compared SCL with a clustering-based representation learning approach [3] and two general instance contrastive learning schemes: Instance-wise learning (MoCo [9]) and local neighbourhood discrimination based learning (PAD [17]). The learned feature representations from both models are applied with K-means for clustering. As shown

Dataset	CIFAR-10	CIFAR-100	STL-10
MoCo [9]*	0.528	0.360	0.561
PAD [17]†	0.626	0.288	0.465
DeepCluster [3]†	0.374	0.189	0.334
SCL	**0.813**	**0.482**	**0.638**

Table 2: Comparisons to representation learning methods. Notation: $(\cdot)^\star$ indicates results reproduced from scratch using the authors' code [9]; $(\cdot)^\dagger$ are from [17].

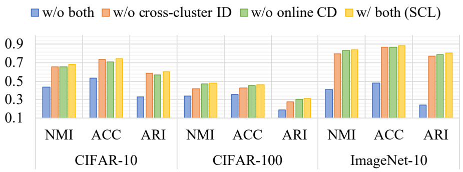


Figure 3: Ablation studies on *cross-cluster* instance descrimination (ID) and *online* cluster discrimination (CD) designs.

in Table 2, our SCL method outperformed all the representation learning methods across the board. This shows clearly the advantages of SCL from holistically modelling the inherent class structure, resulting also a more optimal representation, as compared to separating representation learning from class membership estimation.

4.2 Ablation Study

Detailed ablation studies were conducted for in-depth analysis of SCL. K-means was adopted for models which did not yield desired number of clusters. Experimental results were averaged over multiple trials.

Instance and cluster discrimination. We investigated the independent contributions of our *cross-cluster* instance discrimination (ID) and *online* cluster discrimination (CD) designs in the SCL model. For models trained without cross-cluster ID, all the memory banks were concatenated as the contrastive set for every sample (Eq. (2)), whilst the cluster assignments $\tilde{\boldsymbol{p}}$ yielded by the semantic memory (Eq. (3)) was used for pseudo labels updating if learned without online CD. Instance contrastive learning was considered as the baseline without both the ID and CD components of SCL. As shown in Fig. 3, the models trained without cross-cluster ID or online CD can always surpass instance contrastive learning with remarkable margins, which demonstrates their effectiveness as individual components. By jointly learning with both, SCL always produced superior performances which indicates the mutual benefits of representation learning and decision boundaries reasoning.

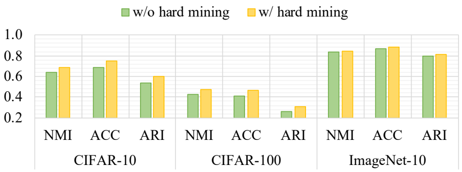


Figure 4: An ablation study on the hard sample mining strategy.

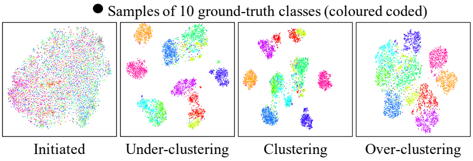
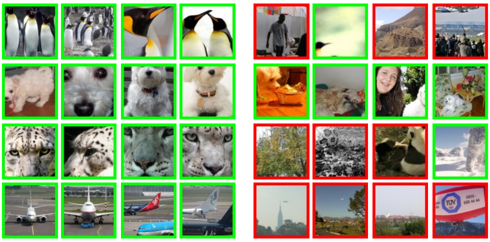


Figure 5: Feature visualisation for images on CIFAR-10 using t-SNE [30].

Hard sample mining. To emphasise the hard samples in model learning, we re-weighted the samples within the same mini-batches according to their assignment stability (Eq. (6)). To study the effectiveness of this design, we replaced it by averaging their losses as in conventional batch-wise training. According to Fig. 4, the learned clusters show higher consistency with the ground-truth classes when training with the re-weighting strategy. This demonstrates the importance of higtlighting hard samples with ambiguous semantic meanings to improve the model's class discrimination capability.

Feature visualisation. To better understand model effectiveness, we visualise some sample representations from a randomly initialised model and those learned from different clustering tasks with different cluster numbers (under-, over- and clustering) on CIFAR-10. Fig. 5 shows that the initial states of the feature spaces were chaotic, which would certainly lead to error-propagation if trained by estimated assignments in a conventional supervised learning process. Whilst the 'over-clustering' task resulted in less within-cluster compactness than 'clustering' and 'under-clustering', 'under-clustering' yielded less separable clusters. By jointly training on all three tasks, SCL explores visual similarity in multiple granularities and learns clusters according to their consensus, hence, more robust to visual ambiguity.

Visual case examples. Fig. 6 shows two groups of image examples from ImageNet-10, with the highest/lowest probabilities (confident/unconfident) for being in a cluster shown in each row. It is evident that the assignment confidence yielded by SCL is well-aligned with the correctness of model predictions. This means that the most confident label interpretations of the learned clusters are also more likely in agreement with the ground-truth categories, *i.e.* semantic plausibility is consistent with the model prediction confidence. Most of the failed cases are due to images being significantly dominated by background. This suggests that it is challenging for unsupervised learning to identify correctly the relevant focus of attention in a visual context.



(a) Confident cases (b) Unconfident cases

Figure 6: Examples from ImageNet-10. Per class each row: (a) top-4 'confident' and (b) bottom-4 'unconfident' cases w.r.t. assignment probabilities. Samples in green boxes are assigned to the correct classes while those with red boxes are failed cases.

5 Conclusion

In this work, we proposed a novel *Semantic Contrastive Learning* (SCL) method for high-level semantic understanding of visual data without learning from manual labels. The SCL model addresses the fundamental limitation of instance contrastive learning by imposing the cluster structure into the unlabelled training data so to jointly learn discriminative visual feature representations and reason about cluster decision boundaries while avoiding the inherent contradiction between their learning objectives. By learning visual features with high robustness to temporal (intermediate) cluster assignments in the course of model training, SCL mitigates the common error-propagation problem of contemporary deep clustering techniques. Moreover, by exploring semantic relations from contrastive visual similarity, the clusters yielded by SCL encode unique and consistent visual characteristics. Hence, SCL is semantically more plausible. Experiments on six object recognition datasets show the SCL's superiority over the state-of-the-art deep clustering and instance contrastive models.

Acknowledgements

This work was supported by the China Scholarship Council, Vision Semantics Limited, the Alan Turing Institute Turing Fellowship.

References

[1] Radhakrishna Achanta and Sabine Susstrunk. Superpixels and polygons using simple non-iterative clustering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4651–4660, 2017.

[2] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *Int. Conf. Learn. Represent.*, 2020.

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Eur. Conf. Comput. Vis.*, pages 1–18, 2018.

[4] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *Int. Conf. Comput. Vis.*, pages 2959–2968, 2019.

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Adv. Neural Inform. Process. Syst.*, volume 33, pages 9912–9924, 2020.

[6] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *Int. Conf. Comput. Vis.*, pages 5879–5887, 2017.

[7] Jianlong Chang, Yiwen Guo, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep discriminative clustering analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1–11, 2019.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[10] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. pages 215–223, 2011.

[11] Kien Do, Truyen Tran, and Svetha Venkatesh. Clustering by maximizing mutual information across views. In *Int. Conf. Comput. Vis.*, pages 9928–9938, 2021.

[12] Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. Improved deep embedded clustering with local structure preservation. In *IJCAI*, pages 1753–1759, 2017.

[13] Philip Haeusser, Johannes Plapp, Vladimir Golkov, Elie Aljalbout, and Daniel Cremers. Associative deep clustering: Training a classification network with no labels. pages 18–32. Springer, 2018.

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9729–9738, 2020.

[15] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE, 2016.

[16] Jiabo Huang, Qi Dong, Shaogang Gong, and Xiatian Zhu. Unsupervised deep learning by neighbourhood discovery. 2019.

[17] Jiabo Huang, Qi Dong, Shaogang Gong, and Xiatian Zhu. Unsupervised deep learning via affinity diffusion. In *AAAI*, 2020.

[18] Pan Ji, Tong Zhang, Hongdong Li, Mathieu Salzmann, and Ian Reid. Deep subspace clustering networks. In *Adv. Neural Inform. Process. Syst.*, pages 24–33, 2017.

[19] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information distillation for unsupervised image segmentation and clustering. In *Int. Conf. Comput. Vis.*, pages 1–10, 2019.

[20] Shaogang Gong Jiabo Huang and Xiatian Zhu. Deep semantic clustering by partition confidence maximisation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[21] Armand Joulin, Francis Bach, and Jean Ponce. Discriminative clustering for image co-segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1943–1950. IEEE, 2010.

[22] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Adv. Neural Inform. Process. Syst.*, 33:18661–18673, 2020.

[23] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[24] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[25] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 2015.

[26] Junnan Li, Pan Zhou, Caiming Xiong, and Steven C.H. Hoi. Prototypical contrastive learning of unsupervised representations. In *ICLR*, 2021.

[27] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *AAAI*, 2021.

[28] Zichuan Liu, Guosheng Lin, Sheng Yang, Jiashi Feng, Weisi Lin, and Wang Ling Goh. Learning markov clustering networks for scene text detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1–11. IEEE, 2018.

[29] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[30] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[31] Chuang Niu, Jun Zhang, Ge Wang, and Jimin Liang. Gatcluster: Self-supervised gaussian-attention network for image clustering. In *Eur. Conf. Comput. Vis.*, 2020.

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In *Adv. Neural Inform. Process. Syst.*, 2018.

[33] Xi Peng, Jiashi Feng, Jiwen Lu, Wei-Yun Yau, and Zhang Yi. Cascade subspace clustering. In *AAAI*, 2017.

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015.

[35] Yaling Tao, Kentaro Takagi, and Kouta Nakata. Clustering-friendly representation learning via instance discrimination and feature decorrelation. In *Int. Conf. Learn. Represent.*, 2021.

[36] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[37] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *Eur. Conf. Comput. Vis.*, pages 268–285. Springer, 2020.

[38] Jianlong Wu, Keyu Long, Fei Wang, Chen Qian, Cheng Li, Zhouchen Lin, and Hongbin Zha. Deep comprehensive correlation mining for image clustering. In *Int. Conf. Comput. Vis.*, pages 1–12, 2019.

[39] Zhirong Wu, Yuanjun Xiong, X Yu Stella, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[40] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. pages 478–487, 2016.

[41] Bo Yang, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. pages 3861–3870. JMLR. org, 2017.

[42] Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5147–5156, 2016.

[43] Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. Online deep clustering for unsupervised representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6688–6697, 2020.

[44] Junjie Zhao, Donghuan Lu, Kai Ma, Yu Zhang, and Yefeng Zheng. Deep image clustering with category-style representation. In *Eur. Conf. Comput. Vis.*, 2020.

[45] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Int. Conf. Comput. Vis.*, pages 6002–6012, 2019.

[46] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991, 2019.