Unsupervised Deep Learning by Neighbourhood Discovery

Jiabo Huang [1] **Qi Dong** [1] **Shaogang Gong** [1] **Xiatian Zhu** [2]

Abstract

Deep convolutional neural networks (CNNs) have demonstrated remarkable success in computer vision by *supervisedly* learning strong visual feature representations. However, training CNNs relies heavily on the availability of exhaustive training data annotations, limiting significantly their deployment and scalability in many application scenarios. In this work, we introduce a generic *unsupervised deep learning* approach to training deep models without the need for any manual label supervision. Specifically, we progressively discover sample anchored/centred neighbourhoods to reason and learn the underlying class decision boundaries iteratively and accumulatively. Every single neighbourhood is specially formulated so that all the member samples can share the same unseen class labels at high probability for facilitating the extraction of class discriminative feature representations during training. Experiments on image classification show the performance advantages of the proposed method over the state-of-the-art unsupervised learning models on six benchmarks including both coarse-grained and fine-grained object image categorisation.

1. Introduction

Deep neural networks, particularly convolutional neural networks (CNNs), have significantly advanced the progress of computer vision problems (Goodfellow et al., 2016; LeCun et al., 2015). However, such achievements are largely established upon *supervised learning* of network models on a massive collection of exhaustively labelled training imagery data (Krizhevsky et al., 2012a; Dong et al., 2019; 2018). This dramatically restricts their scalability and usability to many practical applications with limited labelling budgets. A natural solution is *unsupervised learning* of deep feature representations, which has recently drawn increasing attention (Wu et al., 2018; Caron et al., 2018).

In the literature, representative unsupervised deep learning methods include clustering (Caron et al., 2018; Xie et al., 2016; Yang et al., 2017) and sample specificity analysis (Wu et al., 2018; Bojanowski & Joulin, 2017). The objective of clustering is to identify a set of clusters and each represents an underlying class concept. This strategy has great potential with the best case reaching to the performance of supervised learning, but is error-prone due to the enormous combinatorial space and complex class boundaries. In contrast, sample specificity learning avoids the cluster notion by treating every single sample as an independent class. The hypothesis is that the model can reveal the underlying class-to-class semantic similarity structure, e.g. the manifold geometry. Whilst collecting such instance labels requires no manual annotation cost, the resulting supervision is ambiguous therefore weak to class discrimination. Other contemporary self-supervised learning methods (Doersch et al., 2015; Zhang et al., 2016; Noroozi & Favaro, 2016; Noroozi et al., 2017; Zhang et al., 2017) share a similar limitation due to the insufficient correlation between the auxiliary supervision and the underlying class target.

In this work, we present a generic unsupervised deep learning method called *Anchor Neighbourhood Discovery* (AND). The AND model combines the advantages of both clustering and sample specificity learning whilst mitigating their disadvantages in a principled formulation. Specifically, with a *divide-and-conquer* principle, the AND discovers class consistent neighbourhoods anchored to individual training samples (*divide*) and propagates the local inter-sample class relationships within such neighbourhoods (*conquer*) for more reliably extracting the latent discrimination information during model training. Neighbourhoods can be considered as tiny sample anchored clusters with higher compactness and class consistency. They are specially designed for minimising the clustering errors whilst retaining the exploration of inter-sample class information that is entirely neglected in sample specificity learning. To enhance the neighbourhood quality (class consistency), we introduce a progressive discovery curriculum for incrementally deriving more accurate neighbourhood supervision.

We make three **contributions**: **(1)** We propose the idea

[1] Queen Mary University of London [2] Vision Semantics Limited. Correspondence to: Xiatian Zhu <eddy.zhuxt@gmail.com>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

of exploiting local neighbourhoods for unsupervised deep learning. This strategy preserves the capability of clustering for class boundary inference whilst minimising the negative impact of class inconsistency typically encountered in clusters. To our best knowledge, it is the first attempt at exploring the concept of neighbourhood for end-to-end deep learning of feature representations without class label annotations. **(2)** We formulate an *Anchor Neighbourhood Discovery* (AND) approach to progressive unsupervised deep learning. The AND model not only generalises the idea of sample specificity learning, but also additionally considers the originally missing sample-to-sample correlation during model learning by a novel neighbourhood supervision design. **(3)** We further introduce a curriculum learning algorithm to gradually perform neighbourhood discovery for maximising the class consistency of neighbourhoods therefore enhancing the unsupervised learning capability.

Extensive experiments are conducted on four coarse-grained (CIFAR10 and CIFAR100 (Krizhevsky & Hinton, 2009), SVHN (Netzer et al., 2011), ImageNet (Russakovsky et al., 2015)) and two fine-grained (CUB200-2011 (Wah et al., 2011) and Stanford Dogs (Khosla et al., 2011)) object image classification datasets. The results show the advantages of our AND method over a wide variety of existing state-of-the-art unsupervised deep learning models.

2. Related Work

Existing unsupervised deep learning methods generally fall into four different categories: **(1)** Clustering analysis (Caron et al., 2018; Xie et al., 2016; Yang et al., 2017), **(2)** Sample specificity learning (Wu et al., 2018; Bojanowski & Joulin, 2017), **(3)** Self-supervised learning (Doersch et al., 2015; Zhang et al., 2016; Noroozi & Favaro, 2016; Noroozi et al., 2017; Zhang et al., 2017), and **(4)** Generative models (Goodfellow et al., 2014; Vincent et al., 2010).

Clustering analysis is a long-standing approach to unsupervised machine learning (Aggarwal & Reddy, 2013). With the surge of deep learning techniques, recent studies have attempted to optimise clustering analysis and representation learning jointly for maximising their complementary benefits (Caron et al., 2018; Xie et al., 2016; Yang et al., 2017; Dizaji et al., 2017). Regardless, the key remains the discovery of multiple class consistent clusters (or groups) on the entire training data. This is a difficult task with the complexity and solution space exponentially proportional to both the data and cluster size. It is particularly so for clustering the data in complex structures and distributions such as images and videos. In contrast, the proposed AND model replaces the clustering operation with local neighbourhood identification in a *divide-and-conquer* principle. This enables the control and mitigation of the clustering errors and their negative propagation, potentially yielding more accurate inference of latent class decision boundaries.

Sample specificity learning goes to the other extreme by considering every single sample as an independent class (Wu et al., 2018; Bojanowski & Joulin, 2017). The key idea is that supervised deep learning of neural networks automatically reveals the visual similarity correlation between different classes from end-to-end optimisation. However, this sort of supervision does not explicitly model the class decision boundaries as clustering analysis and the AND model. It is therefore likely to yield more ambiguous class structures and less discriminative feature representations.

Self-supervised learning has recently gained increasing research efforts (Doersch et al., 2015; Zhang et al., 2016; Noroozi & Favaro, 2016; Noroozi et al., 2017; Zhang et al., 2017). Existing methods vary essentially in the design of unsupervised auxiliary supervision. Typically, such auxiliary supervision is hand-crafted to exploit some information intrinsically available in the unlabelled training data, such as spatial context (Doersch et al., 2015; Noroozi & Favaro, 2016), spatio-temporal continuity (Wang & Gupta, 2015; Wang et al., 2017), and colour patterns (Zhang et al., 2016; Larsson et al., 2016). Due to the weak correlation with the underlying class targets, such learning methods mostly yield less discriminative models than clustering analysis and our AND method. How to design more target related auxiliary supervision remains an open problem.

Generative model is a principled way of learning the true data distribution of the training set in an unsupervised manner. The most commonly used and efficient generative models include Restricted Boltzmann Machines (Lee et al., 2009; Hinton et al., 2006; Tang et al., 2012), Autoencoders (Ng, 2011; Vincent et al., 2010), and Generative Adversarial Networks (Radford et al., 2016; Goodfellow et al., 2014). The proposed AND model does not belong to this family, but potentially generates complementary feature representations due to a distinct modelling strategy.

Broadly, AND relates to constrained clustering (Wagstaff et al., 2001; Kamvar et al., 2003; Zhu et al., 2013; 2016) if considering our neighbourhood constraint as a form of pairwise supervision including must-link and cannot-link. However, our method is totally unsupervised without the need for pairwise links therefore more scalable.

3. Unsupervised Neighbourhood Discovery

Suppose we have N training images $\mathcal{I} = \{\boldsymbol{I}_1, \boldsymbol{I}_2, ..., \boldsymbol{I}_N\}$. In unsupervised learning, no class labels are annotated on images. The objective is to derive a deep CNN model $\boldsymbol{\theta}$ from the imagery data $\mathcal{I}$ that allows to extract class discriminative feature representations $\boldsymbol{x}$, $f_{\boldsymbol{\theta}} : \boldsymbol{I} \rightarrow \boldsymbol{x}$. Without the access to class labels, it is unsupervised how the feature points $\boldsymbol{x}$ should be distributed in training so that they can correctly represent the desired class memberships. It is therefore necessary for an unsupervised learning algorithm to reveal such discriminative information directly from the visual data. This is challenging due to the arbitrarily complex appearance patterns and variations typically exhibited in the image collections both within and across classes, implying a high complexity of class decision boundaries.

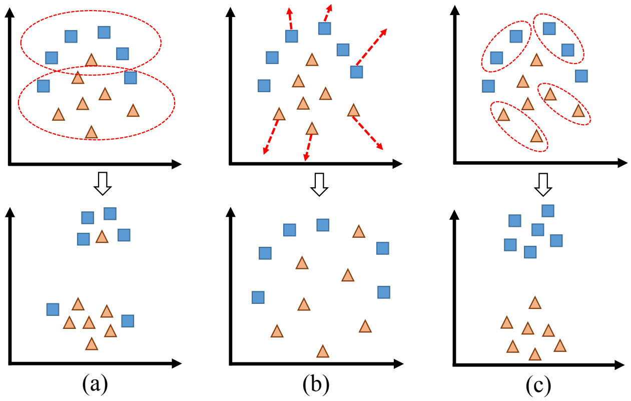


Figure 1. Illustration of three unsupervised learning strategies. **(a)** *Clustering analysis* aims for discovering the global class decision boundary (Caron et al., 2018; Xie et al., 2016); **(b)** *Sample specificity learning* discards the concept of clusters by treating every training sample as an independent class (Wu et al., 2018; Bojanowski & Joulin, 2017); **(c)** Our *Anchor Neighbourhood Discovery* searches local neighbourhoods with high class consistency.

To overcome the aforementioned problem, we formulate an ***Anchor Neighbourhood Discovery*** (AND) method. It takes a *divide-and-conquer* strategy from the local sample anchored neighbourhood perspective. The key idea is that, whilst it is difficult and error-prone to directly reason the *global class decision boundaries* at the absence of class labels on the training data (Fig 1(a)), it would be easier and more reliable to estimate *local class relationship* in small neighbourhoods (Fig 1(c)). Although such information is *incomplete* and provides *less* learning supervision than the conventional clustering strategy (Caron et al., 2018; Xie et al., 2016) that operates at the coarse group level and mines the clusters of data samples, it favourably mitigates the misleading effect of noisy supervision. Besides, the proposed AND model differs dramatically from the sample specificity learning strategy (Wu et al., 2018; Bojanowski & Joulin, 2017) that lacks a fundamental ability to mine the inter-sample class relationships primitive to the global class boundaries (Fig 1(b)). Therefore, the proposed method represents a conceptual trade-off between the two existing strategies and a principled integration of them.

As shown in our evaluations, the proposed training strategy yields superior models. This indicates the significance of both minimising the erroneous self-mined supervision and exploiting the inter-sample class relations spontaneously during unsupervised learning. An overview of the proposed AND model is depicted in Fig 2.

3.1. Neighbourhood Discovery

We start with how to identify neighbourhoods. An intuitive method is using k nearest neighbours (kNN) given a feature space X and a similarity metric s, e.g. the cosine similarity (Fig 2(b)). A neighbourhood $\mathcal{N}_k(\boldsymbol{x})$ determined by kNN is sample-wise, i.e. anchored to a specific training sample $\boldsymbol{x}$:

$$\mathcal{N}_k(\boldsymbol{x}) = \{\boldsymbol{x}_i \mid s(\boldsymbol{x}_i, \boldsymbol{x}) \text{ is top-}k \text{ in } X\} \cup \{\boldsymbol{x}\}, \tag{1}$$

where X denotes the feature space. We call such structures as ***Anchor Neighbourhoods*** (`AN`).

To enable class discriminative learning, we want all samples in a single neighbourhood `AN` to share the same class label, i.e. *class consistent*. As such, we can facilitate the design of learning supervision by assigning the same label to these samples. This requirement, however, is non-trivial to fulfil in unsupervised learning since we have no reasonably good sample features, even though a neighbourhood `AN` can be much smaller and more local (therefore likely more class consistent) than a typical cluster when using small k values. Moreover, we begin with the training images but *no* learned features. This even prevents the formation of $\mathcal{N}_k$ and gives rise to an extreme case – each individual sample represents a distinct anchor neighbourhood.

Neighbourhood Initialisation. Interestingly, such initial `AN`s are in a similar spirit of sample specificity learning (Wu et al., 2018; Bojanowski & Joulin, 2017) where each data instance is assumed to represent a distinct class (Fig 2(a)). With this conceptual linkage, we exploit the instance loss (Wu et al., 2018) to commence the model learning. Specifically, it is a non-parametric variant of the softmax cross-entropy loss written as:

$$\mathcal{L}_{\text{init}} = -\sum_{i=1}^{n_{\text{bs}}} \log(p_{i,i}), \;\; p_{i,j} = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{x}_j/\tau)}{\sum_{k=1}^{N} \exp(\boldsymbol{x}_i^\top \boldsymbol{x}_k/\tau)} \tag{2}$$

where n_{bs} denotes the training mini-batch size, and the temperature parameter τ is for controlling the distribution concentration degree (Hinton et al., 2014).

Neighbourhood Supervision. In the feature space derived by Eq (2), we build a neighbourhood $\mathcal{N}_k(\boldsymbol{x})$ for each individual sample $\boldsymbol{x}$. Considering the high appearance similarity among the samples of each $\mathcal{N}_k(\boldsymbol{x})$, we assume they share a single class label for model discriminative learning.

Formally, we formulate an unsupervised neighbourhood supervision signal as:

$$\mathcal{L}_{\text{AN}} = -\sum_{i=1}^{n_{\text{bs}}} \log\Big(\sum_{j \in \mathcal{N}_k(\boldsymbol{x}_i)} p_{i,j}\Big) \tag{3}$$

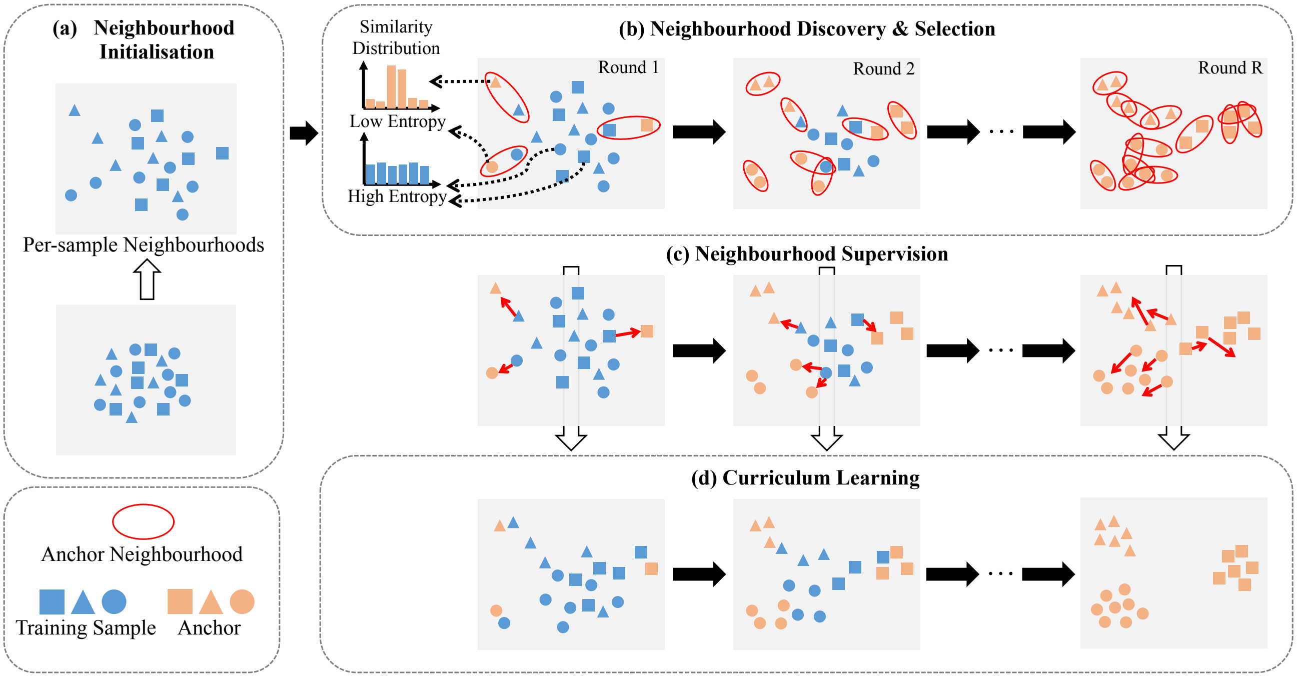


Figure 2. Overview of the proposed *Anchor Neighbourhood Discovery* (AND) method for unsupervised deep learning. **(a)** The AND model starts with per-sample neighbourhoods for model initialisation. **(b)** The resulting feature representations are then used to discover the local neighbourhoods anchored to every single training sample, i.e. anchor neighbourhoods. **(c)** To incorporate the neighbourhood structure information into model learning, we propose a differentiable neighbourhood supervision loss function for enabling end-to-end model optimisation. **(d)** For enhancing model discriminative learning, we further derive a curriculum learning algorithm for selecting class consistent neighbourhoods in a progressive manner. This is based on a novel similarity distribution entropy measurement.

The rationale behind Eq (3) is to encourage label consistency for anchor neighbourhoods (Fig 2(c)). Specifically, the probability $p_{i,j}$ (Eq (2)), obtained using a softmax function, represents visual similarity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in a stochastic manner. This takes the spirit of *stochastic nearest neighbour* (Goldberger et al., 2005), as it considers the entire training set. In this scheme, the probability $p(\boldsymbol{x})$ of correctly classifying a sample $\boldsymbol{x}_i$ can be then represented as:

$$p(\boldsymbol{x}_i) = \sum_{j \in C} p_{i,j} \tag{4}$$

where C denotes the set of samples in the same class as $\boldsymbol{x}_i$. However, C is unavailable to unsupervised learning. To overcome this problem, we approximate C by the neighbourhoods ANs, each of which is likely to be class consistent. Together with the cross-entropy function, this finally leads to the formulation of the proposed $\mathcal{L}_{\text{AN}}$ loss (Eq (3)).

Remarks. The proposed neighbourhood supervision formulation $\mathcal{L}_{\text{AN}}$ aims at exploring the *local class information*, under the assumption that anchor neighbourhoods are class consistent. This is because each neighbourhood AN is treated as a different learning concept (e.g. class), although some ANs may share the same unknown class label. Such information is also *partial* due to that a specific AN may represent only a small proportion of the corresponding class, and multiple ANs with the same underlying class can represent different aspects of the same concept *collectively* (not the whole view due to no AN-to-AN relationships). It is the set of these distributed anchor neighbourhoods *as a whole* that brings about the class discrimination capability during model training. It is in a *divide-and-conquer* principle.

Fundamentally, the proposed design differs dramatically from both (1) *the clustering strategy* that seeks for the complete class boundary information – a highly risky and error-prone process (Caron et al., 2018; Xie et al., 2016), and (2) the *sample specificity learning* that instead totally ignores the class level information therefore less discriminative (Wu et al., 2018; Bojanowski & Joulin, 2017). Moreover, clustering often requires the prior knowledge of the cluster number therefore limiting their usability and scalability due to the lack of it in many applications. On the contrary, this kind of information is not needed for forming the proposed ANs, therefore more application generic and scalable. To maximise the class consistency degree in ANs, we simply need to use the smallest neighbourhood size, i.e. $k=1$.

Neighbourhood Selection. As discussed above, the proposed method requires the neighbourhoods ANs to be class consistent. This condition, nonetheless, is difficult to meet. Specifically, the instance loss function $\mathcal{L}_{\text{init}}$ (Eq (2)) encourages the feature representation learning towards that each sample's specificity degree can be maximised as possible on the training data. Considering a sample $\boldsymbol{x}_i$, other samples either share the class label (positive) with $\boldsymbol{x}_i$ or not (negative). Hence, this formulation may yield a model with certain discrimination ability, e.g. when a subset of (unknown) positive samples are associated with similar visual specificity. But this entirely depends on the intrinsic data properties without stable guarantee. It means that typically *not* all neighbourhoods ANs are reliable and class consistent. This inevitably leads to the necessity of conducting neighbourhood selection for more reliable model learning.

To this end, we go beyond by taking advantages of the curriculum learning idea (Bengio et al., 2009; Dong et al., 2017). Instead of taking a one-off neighbourhood selection, we introduce a *progressive* selection process (Fig 2(d)) which distributes evenly the neighbourhood selection across R rounds. This realises an easy-to-hard learning procedure through a curriculum.

Selecting Curriculum. To enable automated neighbourhood selection for making a scalable curriculum, it is necessary for us to derive a selecting criterion. This is achieved by exploiting the intrinsic nature of the probability $p_{i,j}$ (Eq (2)) defined between two samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. More specifically, we utilise the entropy measurement of the probability vector $p_i = [p_{i,1}, p_{i,2}, \cdots, p_{i,N}]$ as the class consistency indicator of the corresponding neighbourhood AN as:

$$H(\boldsymbol{x}_i) = -\sum_{j=1}^{N} p_{i,j} \log(p_{i,j}). \tag{5}$$

We consider that smaller $H(\boldsymbol{x}_i)$ values correspond to more consistent neighbourhoods. In particular, when $H(\boldsymbol{x}_i)$ is small, it means $\boldsymbol{x}_i$ resides in a low-density area with sparse visual similar neighbours surrounding. In the definition of sample specificity learning (Eq (2)), the model training tends to converge to some local optimum that all samples of a neighbourhood $\mathcal{N}_k(\boldsymbol{x}_i)$ with small $H(\boldsymbol{x}_i)$ share some easy-to-locate visual appearance, and simultaneously the same underlying class label statistically since positive samples are more likely to present such appearance commonness including the context than negative ones. On the contrary, a large $H(\boldsymbol{x}_i)$ implies a neighbourhood $\mathcal{N}_k(\boldsymbol{x}_i)$ residing in a dense area, a case that the model fails to identify the sample specificity. This is considered hard cases, and requires more information for the model to interpret them.

In light of the observations above, we formulate a linear curriculum according to the class consistency entropy measurement. Specifically, for the r-th round (among a total of R rounds), we select the top-S (Eq (6)) of ANs according to their corresponding entropy for model learning by the proposed neighbourhood supervision loss $\mathcal{L}_{\text{AN}}$ (Eq (3)).

$$S = \frac{r}{R} * 100\% \tag{6}$$

Since the remaining training samples are still not sufficiently interpreted by the model at the current round, they are preserved as individuals (i.e. single-sample neighbourhoods) as in sample specificity learning (Eq (2)).

Objective Loss Function. With the progressive neighbourhood discovery as above, we obtain the model objective loss function for the r-th round as:

$$\mathcal{L}^r = -\sum_{i \in B^r_{\text{inst}}} \log(p_{i,i}) - \sum_{i \in B^r_{\text{AN}}} \log\Big(\sum_{j \in \mathcal{N}_k(\boldsymbol{x}_i)} p_{i,j}\Big) \tag{7}$$

where B^r_{inst} and B^r_{AN} denote the set of instances and the set of ANs in a mini-batch at the r-th round, respectively.

As each round of training is supposed to improve the model, we update the neighbourhoods ANs for all training samples before performing neighbourhood selection per round. To facilitate this process, we maintain an offline memory to store the feature vectors. We update the memory features of mini-batch samples by exponential moving average (Lucas & Saccucci, 1990) over the training iterations as:

$$\tilde{\boldsymbol{x}}_i = (1-\eta) \cdot \tilde{\boldsymbol{x}}_i + \eta \cdot \boldsymbol{x}_i \tag{8}$$

where η denotes the update momentum, $\boldsymbol{x}_i$ and $\tilde{\boldsymbol{x}}_i$ the up-to-date and memory feature vector respectively.

3.2. Model Optimisation

The proposed loss function (Eq (7)) is differentiable therefore enabling the stochastic gradient descent algorithm for model training. In particular, when $\boldsymbol{x}_i$ comes as an individual instance, the gradients for $\mathcal{L}^r$ w.r.t. $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ ($j \neq i$) are written as:

$$\frac{\partial \mathcal{L}^r}{\partial \boldsymbol{x}_i} = \frac{1}{\tau}[\sum_{k=1}^{N}(p_{i,k} \cdot \boldsymbol{x}_k) + (p_{i,i} - 2) \cdot \boldsymbol{x}_i], \quad \frac{\partial \mathcal{L}^r}{\partial \boldsymbol{x}_j} = \frac{1}{\tau} p_{i,j} \cdot \boldsymbol{x}_i \tag{9}$$

When $\boldsymbol{x}_i$ corresponds to an AN, the gradients are then:

$$\frac{\partial \mathcal{L}^r}{\partial \boldsymbol{x}_i} = \frac{1}{\tau}[\sum_{k=1}^{N}(p_{i,k} \cdot \boldsymbol{x}_i) - \sum_{k \in \mathcal{N}_k(\boldsymbol{x}_i)} \tilde{p}_{i,k} + (p_{i,i} - \tilde{p}_{i,i}) \cdot \boldsymbol{x}_i] \tag{10}$$

$$\frac{\partial \mathcal{L}^r}{\partial \boldsymbol{x}_j} = \begin{cases} \frac{1}{\tau}[p_{i,j} \cdot \boldsymbol{x}_i - \tilde{p}_{i,j} \cdot \boldsymbol{x}_i], & j \in \mathcal{N}_k(\boldsymbol{x}_i) \\ \frac{1}{\tau}[p_{i,j} \cdot \boldsymbol{x}_i], & j \notin \mathcal{N}_k(\boldsymbol{x}_i) \end{cases} \tag{11}$$

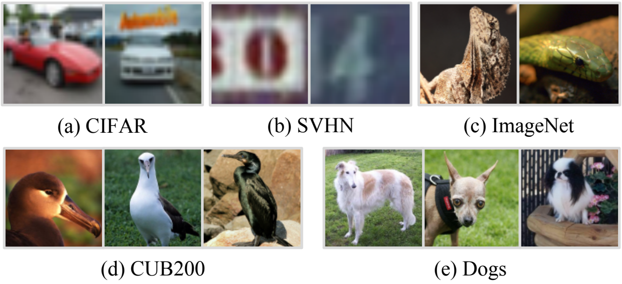
where $\tilde{p}_{i,j} = p_{i,j} / \sum_{k \in \mathcal{N}_k(\boldsymbol{x}_i)} p_{i,k}$ is the normalised distribution over the neighbours. The whole model training procedure is summarised in Algorithm 1.

4. Experiments

Datasets. We used 6 image classification benchmarks for evaluating our model (Fig 3). ***CIFAR10(/100)*** (Krizhevsky & Hinton, 2009): An image dataset with 50,000/10,000

Algorithm 1 Neighbourhood discovery.

```
Input: Training data $\mathcal{I}$, rounds $R$, iterations per round $T$.
Output: A deep CNN feature model.
Initialisation: Instance specificity learning (Eq (2)).
Unsupervised learning:
for $r = 1$ to $R$ do
  Form neighbourhoods with the current features (Eq (1));
  Curriculum selection of neighbourhoods (Eq (6));
  for $t = 1$ to $T$ do
    Network forward propagation (batch feed-forward);
    Objective loss computation (Eq (7));
    Network back-propagation (Eq (9),(10),(11));
    Memory feature update (Eq (8)).
  end for
end for
```



(a) CIFAR (b) SVHN (c) ImageNet

(d) CUB200 (e) Dogs

Figure 3. Dataset example images.

train/test images from 10 (/100) object classes. Each class has 6,000 (/600) images with size 32×32. ***SVHN*** (Netzer et al., 2011): A Street View House Numbers dataset including 10 classes of digit images. ***ImageNet*** (Russakovsky et al., 2015): A large 1,000 classes object dataset with 1.2 million images for training and 50,000 for test. ***CUB200-2011*** (Wah et al., 2011): A fine-grained dataset containing 5,994/5,794 train/test images of 200 bird species. ***Stanford Dogs*** (Khosla et al., 2011): A fine-grained dataset with 12,000/8,580 train/test images of 120 dog breeds.

Experimental setup. For learning any unsupervised representation model, we assumed and used only the training image data but *no* class labels. Unless stated otherwise, we adopted the AlexNet (Krizhevsky et al., 2012b) as the neural network architecture for fair comparisons with the state-of-the-art methods. To assess the quality of a learned model for classification at test time, we utilised the ground-truth class labels of the training images *merely* for enabling image categorisation. This does not change the feature representations derived in unsupervised learning.

Following Wu et al. (2018), we considered two classification models, *Linear Classifier* (LC), and *Weighted kNN*, as well as the feature representations extracted from different network layers respectively. LC was realised by a fully connected (FC) layer optimised by the cross-entropy loss function. The non-parametric classifier kNN predicts the class label by weighted voting of top-k neighbours $\mathcal{N}_k$ as $s_c = \sum_{i \in \mathcal{N}_k} \delta(c, c_i) \cdot w_i$ where $\delta(c, c_i)$ is the Dirac delta function which returns 1 if $c = c_i$, and 0 otherwise. The weight w_i is computed from the cosine similarity s_i as $w_i = \exp(s_i/\tau)$ with $\tau = 0.07$ the temperature parameter. Without an extra classifier learning post-process involved, kNN reflects *directly* the discriminative capability of the learned feature representations.

Performance metric. We adopted the top-1 classification accuracy for the model performance measurement.

Competitors. We compared the proposed AND model with four types of state-of-the-art unsupervised learning methods: **(1)** *Generative model*: BiGAN (Donahue et al., 2016); **(2)** *Clustering method*: DeepCluster (Caron et al., 2018); **(3)** *Self-supervised learning methods*: Context (Doersch et al., 2015), Colour (Zhang et al., 2016), Jigsaw (Noroozi & Favaro, 2016), Counting (Noroozi et al., 2017), and Split-Brain (Zhang et al., 2017); **(4)** *Sample specificity learning methods*: Instance (Wu et al., 2018) and Noise As Targets (NAT) (Bojanowski & Joulin, 2017); in total 9 methods.

Implementation details. For fair comparisons, we used the same experimental setting as (Wu et al., 2018; Bojanowski & Joulin, 2017). To train AND models, we set the learning rate to 0.03 which was further scaled down by 0.1 every 40 epochs after the first 80 epochs. We used the batch size of 256 for ImageNet and 128 for others. We set the epoch to 200 per round. We fixed the feature length to 128. We applied the SGD with Nesterov momentum at 0.9. Our model usually converges with $R = 4$ rounds. We set $\eta = 0.5$ in Eq (8) for feature update. We set $k = 1$ (Eq (1)) for exploring the most local neighbourhoods.

4.1. Comparisons to the State-of-the-Art Methods

Small scale evaluation. Table 1 compares the object image classification results on three benchmarks between AND and four unsupervised learning methods. We tested two classification models, weighted kNN using FC features and linear regression using conv5 features. We have these observations: **(1)** The AND method performs best often by large margins over all competitors, except linear regression on CIFAR10 with DeepCluster outperforms marginally. This suggests the performance superiority of our neighbourhood discovery over alternative methods for unsupervised representation learning. **(2)** The margins obtained by kNN tend to be larger than those by linear regression. This indicates that AND features are favourably more ready for *direct* use without extra classifier training as post-processing.

Large scale evaluation. We evaluated the scalability of our AND model using ImageNet. Table 2 compares AND with nine alternative methods. Following the previous studies, we tested all conv layers. The results show that: **(1)** All unsuper-

Dataset	CIFAR10	CIFAR100	SVHN
Classifier/Feature	Weighted kNN / FC		
Split-Brain	11.7	1.3	19.7
Counting	41.7	15.9	43.4
DeepCluster	62.3	22.7	84.9
Instance	60.3	32.7	79.8
AND	**74.8**	**41.5**	**90.9**
Classifier/Feature	Linear Classifier / conv5		
Split-Brain	67.1	39.0	77.3
Counting	50.9	18.2	63.4
DeepCluster	**77.9**	41.9	92.0
Instance	70.1	39.4	89.3
AND	77.6	**47.9**	**93.7**

Table 1. Evaluation on small scale image datasets.

vised learning methods clearly surpass the random features, suggesting their modelling effectiveness consistently. **(2)** AND outperforms all competitors but by smaller margins. This is likely due to using over tiny neighbourhoods (sized 2) for being consistent with small scale datasets. The amount of inter-sample relationships is quadratic to the data size, so bigger neighbourhoods may be beneficial for large datasets in capturing structural information. **(3)** Most unsupervised learning methods yield the respective best representation at intermediate layers when using linear classifier. The plausible reason is that their supervision singles are less correlated with the ground-truth targets.

Classifier	Linear Classifier						kNN
Feature	conv1	conv2	conv3	conv4	conv5	FC	FC
Random	11.6	17.1	16.9	16.3	14.1	12.0	3.5
Supervised	19.3	36.3	44.2	48.3	50.5	-	-
Context	17.5	23.0	24.5	23.2	20.6	30.4	-
BiGAN	17.7	24.5	31.0	29.9	28.0	32.2	-
Colour	13.1	24.8	31.0	32.6	31.8	35.2	-
Jigsaw	**19.2**	30.1	34.7	33.9	28.3	**38.1**	-
NAT	-	-	-	-	-	36.0	-
Counting	18.0	30.6	34.3	32.5	25.7	-	-
Split-Brain	17.7	29.3	35.4	35.2	32.8	-	11.8
DeepCluster	13.4	**32.3**	**41.0**	39.6	**38.2**	-	26.8
Instance	16.8	26.5	31.8	34.1	35.6	-	**31.3**
AND	15.6	27.0	35.9	**39.7**	37.9	36.7	**31.3**

Table 2. Evaluation on ImageNet. The results of existing methods are adopted from (Wu et al., 2018; Bojanowski & Joulin, 2017).

Fine-grained evaluation. We evaluated AND with more challenging fine-grained recognition tasks which are significantly under-studied in unsupervised learning context. Consistent with the results discussed above, Table 3 demonstrates again the performance superiority of our neighbourhood discovery idea over the best competitor Instance.

4.2. Component Analysis and Discussions

We conducted detailed component analysis with the weighted kNN classifier and FC features.

Dataset	CUB200	Dogs
Instance	11.6	27.0
AND	**14.4**	**32.3**

Table 3. Evaluation on fine-grained datasets. Network: ResNet18.

Backbone network. We tested the generalisation of AND with varying-capacity networks. We further evaluated ResNet18 and ResNet101 (He et al., 2016) on CIFAR10. Table 4 shows that AND benefits from stronger net architectures. A similar observation was made on ImageNet: 41.2% (ResNet18) vs. 31.3% (AlexNet).

Network	AlexNet	ResNet18	ResNet101
Accuracy	74.8	86.3	**88.4**

Table 4. Network generalisation analysis on CIFAR10.

Model initialisation. We tested the impact of initial features for neighbourhood discovery by comparing random and Instance networks. Table 5 shows that AND can benefit from stronger initialisation whilst being robust to weak initial representations.

Initialisation	Random Model	Instance Model
Accuracy	85.7	**86.3**

Table 5. Effect of model initialisation on CIFAR10.

Neighbourhood size. Neighbourhood size is an important parameter since it controls label consistency of ANs and finally the model performance. We evaluated its effect using ResNet18 on CIFAR10 by varying k from 1 (the default value) to 100. Figure 4 shows that the smallest neighbourhoods (i.e. $k = 1$) are the best choice. This implies high variety of imagery data, so smaller ANs are preferred for unsupervised learning.

Curriculum round. We tested the effect of the curriculum round (R in Eq (6)) of progressive neighbourhood discovery. More rounds consume higher training costs. Figure 5 shows that using 4 rounds gives a good trade-off between model training efficiency and feature performance. Often, per-round epoch number N_{ep} affects the training efficiency and performance. We investigated its effect and found that AND achieves 83.3% by N_{ep}=50 vs. 84.8% by N_{ep}=100.

One-off *vs.* curriculum neighbourhood discovery. We evaluated the benefit of AND's curriculum. To this end, we compared with the *one-off* discovery counterpart where all anchor neighbourhoods are exploited one time. Table 6 shows that the proposed multi-round progressive discovery via a curriculum is effective to discover more reliable anchor neighbourhoods for superior unsupervised learning.

Neighbourhood quality. We examined the class consistency of anchor neighbourhood discovered throughout the

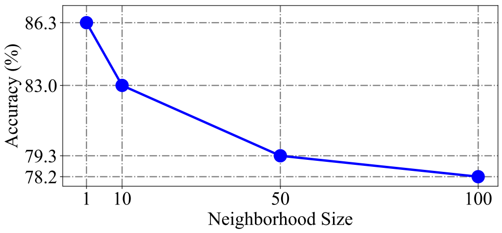


Figure 4. Effect of the neighbourhood size on CIFAR10.

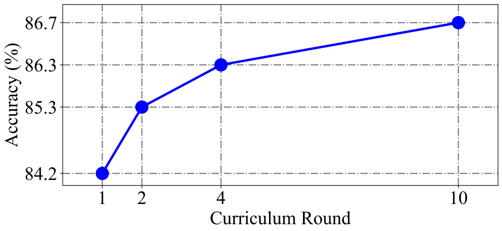


Figure 5. Effect of the curriculum round on CIFAR10.

curriculum rounds. Figure 6 shows that the numbers of both class consistent and inconsistent anchor neighbourhoods increase along with the training rounds, and more importantly the consistent ones raise much more rapidly. This explains the performance advantages of the AND model and the benefit of exploring progressive curriculum learning.

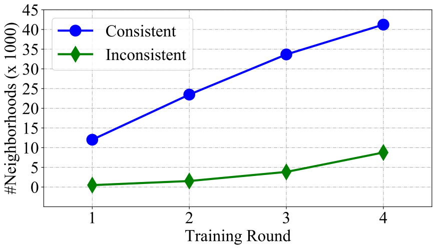


Figure 6. Neighbourhood quality over rounds on CIFAR10.

Learning attention dynamics. To further understand how the AND benefits the feature representation learning, we tracked the modelling attention by Grad-Cam (Selvaraju et al., 2017) to visualise which parts of training images the model is focusing on throughout the curriculum rounds. We have the following observations from Fig 7: **(1)** Often the model initially looks at class irrelevant image regions. This suggests that sample specificity is a less effective supervision signal for guiding model discriminative training. **(2)** In cases, the AND model is able to gradually shift the learning attention towards the class relevant parts therefore yielding a more discriminative model. **(3)** The AND may fail to capture the object attention, e.g. due to cluttered background and poor lighting condition.

Discovery	One-off	Curriculum
Accuracy	84.2	**86.3**

Table 6. One-off vs. curriculum discovery on CIFAR10.

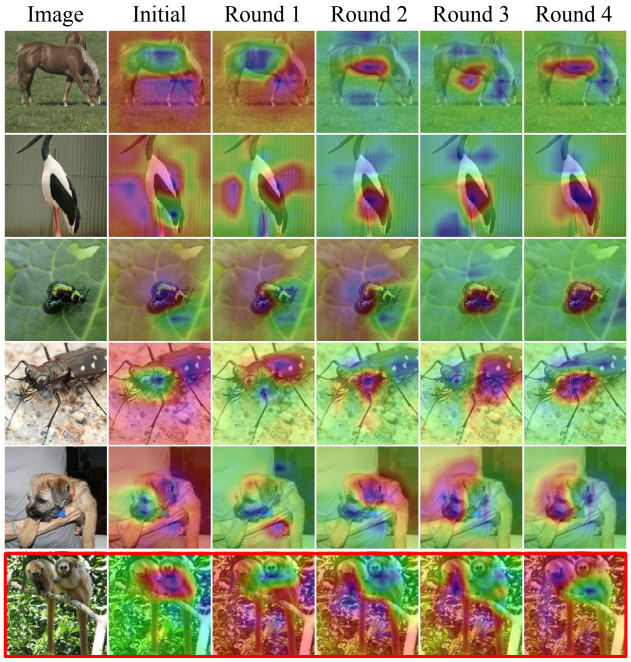


Figure 7. The evolving dynamics of model learning attention throughout the training rounds on six ImageNet classes. Red bounding box indicates a failure case.

5. Conclusion

In this work, we presented a novel Anchor Neighbourhood Discovery (AND) approach for unsupervised learning of discriminative deep network models through class consistent neighbourhood discovery and supervision in a progressive manner. With the AND model, we avoid the notorious grouping noises whilst still preserving the intrinsic merits of clustering for effective inference of the latent class decision boundaries. Our method is also superior to the existing sample specificity learning strategy, due to the unique capability of propagating the self-discovered sample-to-sample class relationship information in end-to-end model optimisation. Extensive experiments on four image classification benchmarks show the modelling and performance superiority of the proposed AND method over a wide range of state-of-the-art unsupervised deep learning models. We also provided in-depth component analysis to give insights on the model advantages of the AND formulation.

Acknowledgements

This work was partly supported by the China Scholarship Council, Vision Semantics Limited, the Royal Society Newton Advanced Fellowship Programme (NA150459), and Innovate UK Industrial Challenge Project on Developing and Commercialising Intelligent Video Analytics Solutions for Public Safety (98111-571149).

References

Aggarwal, C. C. and Reddy, C. K. *Data clustering: algorithms and applications*. CRC press, 2013.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *International Conference on machine learning (ICML)*, pp. 41–48, 2009.

Bojanowski, P. and Joulin, A. Unsupervised learning by predicting noise. In *International Conference on machine learning (ICML)*, pp. 1–10, 2017.

Caron, M., Bojanowski, P., Joulin, A., and Douze, M. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision (ECCV)*, pp. 1–18, 2018.

Dizaji, K. G., Herandi, A., Deng, C., Cai, W., and Huang, H. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 5747–5756, 2017.

Doersch, C., Gupta, A., and Efros, A. A. Unsupervised visual representation learning by context prediction. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 1422–1430, 2015.

Donahue, J., Krähenbühl, P., and Darrell, T. Adversarial feature learning. In *International Conference on Learning Representations (ICLR)*, pp. 1–18, 2016.

Dong, Q., Gong, S., and Zhu, X. Multi-task curriculum transfer deep learning of clothing attributes. In *WACV*, 2017.

Dong, Q., Gong, S., and Zhu, X. Imbalanced deep learning by minority class incremental rectification. *TPAMI*, 2018.

Dong, Q., Zhu, X., and Gong, S. Single-label multi-class image classification by deep logistic regression. *AAAI*, 2019.

Goldberger, J., Hinton, G. E., Roweis, S. T., and Salakhutdinov, R. R. Neighbourhood components analysis. In *Advances in neural information processing systems (NIPS)*, pp. 513–520, 2005.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems (NIPS)*, pp. 2672–2680, 2014.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press Cambridge, 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on computer vision and pattern recognition (CVPR)*, pp. 770–778, 2016.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *Advances in neural information processing systems (NIPS)*, 2014.

Hinton, G. E., Osindero, S., and Teh, Y.-W. A fast learning algorithm for deep belief nets. *Neural computation*, 18 (7):1527–1554, 2006.

Kamvar, K., Sepandar, S., Klein, K., Dan, D., Manning, M., and Christopher, C. Spectral learning. In *International Joint Conference of Artificial Intelligence (IJCAI)*. Stanford InfoLab, 2003.

Khosla, A., Jayadevaprakash, N., Yao, B., and Fei-Fei, L. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012a.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*, pp. 1097–1105, 2012b.

Larsson, G., Maire, M., and Shakhnarovich, G. Learning representations for automatic colorization. In *European Conference on Computer Vision (ECCV)*, pp. 577–593, 2016.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436, 2015.

Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *International Conference on Machine Learning (ICML)*, pp. 609–616, 2009.

Lucas, J. M. and Saccucci, M. S. Exponentially weighted moving average control schemes: properties and enhancements. *Technometrics*, 32(1):1–12, 1990.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep* *learning and unsupervised feature learning*, number 2, pp. 5, 2011.

Ng, A. Sparse autoencoder. pp. 1–19, 2011.

Noroozi, M. and Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision (ECCV)*, pp. 69–84, 2016.

Noroozi, M., Pirsiavash, H., and Favaro, P. Representation learning by learning to count. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 1–9, 2017.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2016.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 1–9, 2017.

Tang, Y., Salakhutdinov, R., and Hinton, G. Robust boltzmann machines for recognition and denoising. In *IEEE Conference on computer vision and pattern recognition (CVPR)*, pp. 2264–2271, 2012.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research (JMLR)*, 11(Dec):3371–3408, 2010.

Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S., et al. Constrained k-means clustering with background knowledge. In *International Conference on Machine Learning (ICML)*, volume 1, pp. 577–584, 2001.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011.

Wang, X. and Gupta, A. Unsupervised learning of visual representations using videos. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 2794–2802, 2015.

Wang, X., He, K., and Gupta, A. Transitive invariance for self-supervised visual representation learning. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 1329–1338, 2017.

Wu, Z., Xiong, Y., Stella, X. Y., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Xie, J., Girshick, R., and Farhadi, A. Unsupervised deep embedding for clustering analysis. In *International Conference on machine learning (ICML)*, pp. 478–487, 2016.

Yang, B., Fu, X., Sidiropoulos, N. D., and Hong, M. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *International Conference on Machine Learning (ICML)*, pp. 1–14, 2017.

Zhang, R., Isola, P., and Efros, A. A. Colorful image colorization. In *European Conference on Computer Vision (ECCV)*, pp. 649–666, 2016.

Zhang, R., Isola, P., and Efros, A. A. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *IEEE Conference on computer vision and pattern recognition (CVPR)*, pp. 1–11, 2017.

Zhu, X., Loy, C. C., and Gong, S. Constrained clustering: Effective constraint propagation with imperfect oracles. In *International Conference on Data Mining (ICDM)*, pp. 1307–1312, 2013.

Zhu, X., Loy, C. C., and Gong, S. Constrained clustering with imperfect oracles. *IEEE transactions on neural networks and learning systems (TNNLS)*, 27(6):1345–1357, 2016.