Qing, Y., Zeng, Y., Cao, Q. and Huang, G.-B. (2021) End-to-end novel visual categories learning via auxiliary self-supervision. *Neural Networks*, 139, pp. 24-32. (doi: [10.1016/j.neunet.2021.02.015](https://doi.org/10.1016/j.neunet.2021.02.015))

The material cannot be used for any other purpose without further permission of the publisher and is for private use only.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

<http://eprints.gla.ac.uk/235959/>

Deposited on 09 March 2021

Enlighten – Research publications by members of the University of
Glasgow

<http://eprints.gla.ac.uk>

End-to-end Novel Visual Categories Learning via Auxiliary Self-Supervision

Yuanyuan Qing^{a,*}, Yijie Zeng^a, Qi Cao^b, Guang-Bin Huang^a

^a*School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798*

^b*School of Computing Science, University of Glasgow, Singapore 56773*

Abstract

Semi-supervised learning has largely alleviated the strong demand for large amount of annotations in deep learning. However, most of the methods have adopted a common assumption that there is always labeled data from the same class of unlabeled data, which is impractical and restricted for real-world applications. In this research work, our focus is on semi-supervised learning when the categories of unlabeled data and labeled data are disjoint from each other. The main challenge is how to effectively leverage knowledge in labeled data to unlabeled data when they are independent from each other, and not belonging to the same categories. Previous state-of-the-art methods have proposed to construct pairwise similarity pseudo labels as supervising signals. However, two issues commonly inherent in these methods: 1) All of previous methods are comprised of multiple training phases, which makes it difficult to train the model in an end-to-end fashion. 2) Strong dependence on the quality of pairwise similarity pseudo labels limits the performance as pseudo labels are vulnerable to noise and bias. Therefore, we propose to exploit the use of self-supervision as auxiliary task during model training such that labeled data and unlabeled data will share the same set of surrogate labels and overall supervising signals can have strong regularization. By doing so, all modules in the proposed algo-

*Corresponding author

Email addresses: qing0006@e.ntu.edu.sg (Yuanyuan Qing), yzeng004@e.ntu.edu.sg (Yijie Zeng), qi.cao@glasgow.ac.uk (Qi Cao), egbhuang@ntu.edu.sg (Guang-Bin Huang)

rithm can be trained simultaneously, which will boost the learning capability as end-to-end learning can be achieved. Moreover, we propose to utilize local structure information in feature space during pairwise pseudo label construction, as local properties are more robust to noise. Extensive experiments have been conducted on three frequently used visual datasets, i.e., CIFAR-10, CIFAR-100 and SVHN, in this paper. Experiment results have indicated the effectiveness of our proposed algorithm as we have achieved new state-of-the-art performance for novel visual categories learning for these three datasets.

Keywords: Image Classification, Novel Visual Categories,
Self-Supervision, Pairwise Similarity, Local Data Structure

1. Introduction

Deep learning has achieved great success on visual recognition in recent years. However, such superior performance heavily relies on large amount of annotated data, which can be expensive, time-consuming and impractical for
5 real-world applications [1]. Semi-supervised learning [2] is utilized to alleviate the strong dependence of data annotation by exploiting structure information in unlabeled data. Many state-of-the-art semi-supervised learning methods are capable of achieving good performance with much less labels than those of supervised baseline [3, 4, 5, 6, 7, 8, 9, 10, 11]. Under a common assumption adopted
10 by semi-supervised learning algorithm, there is always labeled data available from the same class of unlabeled data. However, this assumption cannot be always satisfied especially in real-world situations. More practical scenarios are that there are distinguishing categories for unlabeled data, which means unlabeled data are from separate classes without overlapping with existing classes
15 of labeled data. This notable problem in most semi-supervised learning methods is not well addressed with very few research works on solutions reported in literature. In this research, we focus on semi-supervised learning when the categories of unlabeled data and labeled data are disjoint from each other. If labeled and unlabeled data belong to the same set of categories, label informa-

tion can be efficiently extracted to supervise the model to learn discriminative
 features. On the contrary, label information from labeled data is not directly
 related to unlabeled data if they belong to disjoint categories, which suggests
 that label information is not sufficient as supervising signals. The next research
 question is what else can play the role of supervision and simultaneously apply
 to all data. Several research works [12, 13, 14] have been done recently for novel
 categories learning and their answer to the preceding question is the pairwise
 similarity information. It requires weaker supervision for predicting whether
 two images are similar/belonging to the same class or not, compared to that of
 classifying images into multiple categories. Pairwise similarity pseudo labels are
 employed to supervise model training and the quality of pseudo labels deter-
 mines the ultimate performance of the trained model. [12, 13] train a similarity
 prediction network (SPN) with labeled data first and then the predictions of
 trained SPN on unlabeled data will be used as pairwise pseudo labels. In [14],
 pseudo labels are generated with rank statistics of feature vectors that get up-
 dated during the training process. All of those previous methods are comprised
 of multiple training stages, making it difficult to implement in an end-to-end
 fashion. Moreover, their performances are constrained by strong dependence on
 the quality of pairwise similarity pseudo labels that can be noisy and biased. In
 this paper, we propose an end-to-end novel visual categories learning algorithm
 by introducing a new supervising signal: self-supervision.

Self-supervised learning has demonstrated very prominent results in the field of
 unsupervised feature learning [15]. Pretext tasks are defined in self-supervised
 learning and the required supervising signals are free to get, e.g., angle of ro-
 tational transformation that has been applied [16], color of the image [17] and
 order of patches from the same image [18]. Such self-supervision signals apply
 to both labeled and unlabeled data, and are argued to be able to teach models
 on studying high-level features. Instead of utilizing self-supervision as a pre-
 training stage, which is commonly adopted by most semi-supervised learning
 algorithms [16, 14, 19], self-supervision signals are incorporated with pairwise
 similarity information to supervise the model simultaneously in our proposed

algorithm. Thus, supervising signals injected to the model will have strong regularization so that performance degradation caused by noisy and biased pairwise pseudo labels can be largely alleviated.

Moreover, we propose to further improve the quality of pairwise pseudo labels by exploiting local data structure information in feature space. Methods
55 proposed in [12] and [13] generate pseudo labels with static function/network, whereas [14] constructs pseudo labels on-the-fly with rank statistics. The superior performance of the latter one in [14] inspires us to adopt similar philosophy, i.e., producing pairwise pseudo labels by comparing feature vectors. However,
60 feature space trained without ground-truth label information is noisy. Vanilla similarity metrics, e.g., cosine similarity, may fail to capture correct relationship between feature vectors. Therefore, we propose to use conditional probability in symmetric Stochastic Neighbor Embedding (SNE) [20] as pairwise similarity measure to enforce local properties, which are more robust for noisy feature
65 space.

Our main contributions in this paper are summarized as follows:

- An end-to-end novel visual categories learning algorithm is proposed by utilizing self-supervision signals simultaneously with pairwise similarity information. To the best of our knowledge, it is the first research work
70 that does not require model pre-training such that end-to-end learning can be achieved for novel visual categories learning.
- We propose to utilize robust local structure properties in noisy feature space for the construction of pairwise similarity pseudo labels. In our ablation study, experiment results have verified that pairwise similarity
75 pseudo labels constructed by our algorithm can capture data relationship more accurately than that of commonly-used cosine similarity method.
- Extensive experiments conducted for three commonly-used visual datasets, i.e., CIFAR-10, CIFAR-100 and SVHN, have demonstrated the effectiveness of our proposed algorithm as it is observed that our algorithm has
80 outperformed other state-of-the-art methods on all these three datasets.

The remainder of this paper is organized as follows. Section 2 gives brief introductions of relative works and background knowledge. Detailed explanations of our proposed algorithm are described in Section 3. Experiment configurations and discussions for experiment results are presented in Section 4. Section 5
85 concludes this paper.

2. Relative Works

As our work is closely relevant to semi-supervised learning, self-supervised learning and clustering, we will briefly introduce these three topics in this section.

Semi-supervised learning leverages data structure information in large amount of unlabeled data to improve generalization capability of the model, which would be highly possible to suffer from overfitting issues with limited number of labeled data. Most semi-supervised learning methods can be summarized by Eq. 1:

$$Loss = L_{X_l} + L_{X_u} \quad (1)$$

where L_{X_l} and L_{X_u} denote the loss on labeled data X_l and unlabeled data X_u respectively. For labeled data, the loss term is normally calculated by cross-entropy loss between network prediction and ground-truth label. For unlabeled
90 data, the loss term varies with different methods. Accordingly, semi-supervised learning methods can be classified into two categories: consistency based and pseudo-label based. Consistency based semi-supervised learning is one popular direction, which has shown very encouraging empirical results in many research works [10, 7, 8, 21, 22, 23]. The central strategy is to impose consistency con-
95 straints on network predictions for unlabeled data under perturbations. Such perturbations can be added to data or network. On the other hand, pseudo-label based semi-supervised learning [11, 24, 6] adopts a more straightforward strategy, i.e., to create pseudo labels for unlabeled data and treat pseudo labels as ground-truth labels in subsequent training. [24, 6] have further shown that con-

100 consistency constraints and pseudo-label based methods are complementary with
each other. In our proposed algorithm, we incorporate both consistency-based
and pseudo-label based strategies into model training.

Self-supervised learning is getting increasingly favorable for unsupervised
105 feature representation due to its simplicity and high efficiency. Pretext task
is defined to give model surrogate supervision and methods vary based on the
pretext task defined. RotNet proposed in [16] learns features via predicting the
rotational transformation that has been applied to the image. Jigsaw puzzle
solver proposed in [18] targets to predict relative spatial position of each puzzle
110 tile extracted from the original image. [17] focuses on colorization, i.e., to pre-
dict the color with intensity information of the original image. Self-supervised
learning is commonly used as a pre-training process and the trained model will
be further utilized as a feature extractor for downstream tasks. [14] has also
explored the use of self-supervised learning in novel categories learning by initial-
115 izing the model with trained RotNet [16]. However, self-supervision signals are
utilized independently during pre-training and most blocks of the trained model
are frozen in subsequent training process to avoid overfitting in [14], which limits
the learning capability of the model. Conversely, our proposed algorithm is to
train the model end-to-end by using self-supervision signals with other super-
120 visions jointly, which provides strong regularization along the training process.
As such, the whole network can be trained without concerns on overfitting issues.

Clustering as a classic unsupervised learning problem has been studied in
many research works [25, 26, 27, 28, 29, 30, 31, 32]. In recent years, deep learn-
125 ing based clustering studies [29, 30, 33, 32, 34, 35] have been conducted due to
favorable feature extraction capability of neural networks. Most methods in this
category can simultaneously learn feature embeddings and cluster assignments.
In [29], the model is first pre-trained with deep autoencoder [36] and then op-
timized with pseudo targets computed with current soft cluster assignments.
130 [30] reduces the clustering problem into binary pairwise-classification problem

and similar to [29], pairwise pseudo labels are calculated with current network predictions. However, the criteria used for clustering are ambiguous due to the lack of prior knowledge. Novel categories learning is free from such dilemma as some labeled data is given, which is able to provide categorical information.

135

Zero-shot learning [37, 38, 39, 40, 41, 42, 43] also deals with classification problem of images from unseen categories during training. Unlike common classification problem, extra information is required in zero-shot learning other than training data and testing data, i.e., side information. Class attributes information is one popular option in most zero-shot learning algorithms [44, 38, 45, 46].
 140 The class attributes include high-level semantic information, e.g., color, shape and texture, for each class in training data and testing data. Another form of side information exploited by previous researchers is textual information [47, 48, 49], which can be extracted from Wikipedia text via Word2Vec [50].
 145 Novel categories learning problem discussed in this paper is different from zero-shot learning as above-mentioned additional information is not required in novel categories learning.

3. Methodology

3.1. Problem Formulation

150 Given a set of image data $X = \{x_1, x_2, \dots, x_N\}$, the corresponding labels are only known for part of the classes, denoted by $\{X_l, Y_l\}$, and $Y_l = \{y_1, y_2, \dots, y_l\}$. The number of all possible classes in Y_l is denoted as c_l . For remaining data in X , the corresponding labels are unknown, denoted by $\{X_u, Y_u\}$. The number of all possible classes in Y_u is denoted as c_u , which is assumed to be known.
 155 $Y_u \cap Y_l = \emptyset$, i.e., the label classes in X_u are disjoint from label classes in X_l . The objective is to correctly cluster X_u into the number of c_u classes as much as possible with given $\{X_l, Y_l\}$.

3.2. Proposed Algorithm

The illustration of our proposed algorithm is shown in Fig. 1. Predicting the angle of rotational transformation that has been applied to original images proposed in [16] is utilized in our algorithm as auxiliary self-supervision task. Three different linear classifiers are used to map data points from sharing feature space to three target spaces, i.e., pairwise similarity pseudo label space, categorical label space and self-supervision label space respectively, which correspond to three modules in our algorithm. In the following subsections, detailed explanations are presented for each module.

Figure 1: The illustration of our proposed end-to-end novel visual categories learning algorithm. For both labeled data X_l and unlabeled data X_u , we first rotated the images by $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$. All rotated images are fed into the convolutional neural network to get the corresponding feature vectors ϕ . The feature vectors for 0° rotated images, i.e., the original images, are used to calculate two losses: binary cross-entropy loss $\mathcal{L}_{BCE}$ for X_u and cross-entropy loss $\mathcal{L}_{CE}$ for both X_l and X_u . Feature vectors for all $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$ rotated images are used to calculate the third loss: rotation loss $\mathcal{L}_{Rot}$. f_{c1} , f_{c2} and f_{c3} represent three different linear classifiers (i.e., fully-connected layers), that are employed to map feature vectors to the target space for different loss calculations respectively.

3.2.1. Pairwise Similarity Learning

Without any pre-training or fine-tuning being performed, our algorithm starts with a random initialized CNN network. Image data in X is fed into the network and the corresponding feature vector ϕ_i before linear layers for each $x_i \in X_u$ is collected to compute the pairwise similarity. As there is no label information for X_u , the learned feature space is vulnerable to noise. Therefore, we propose to exploit robust local data structure information in the feature space to mitigate such impact. We use conditional probability proposed in symmetric SNE [20] to represent the similarity s_{ij} between ϕ_i and ϕ_j , computed with Eq. 2 and Eq. 3.

$$p_{i|j} = \frac{\exp(-\|\phi_j - \phi_i\|^2/T^2)}{\sum_{k \neq j} \exp(-\|\phi_j - \phi_k\|^2/T^2)} \quad (2)$$

$$s_{ij} = \frac{1}{2}(p_{i|j} + p_{j|i}) \quad (3)$$

where $p_{i|j}$ represents the conditional probability of ϕ_i being selected as a neighbor of ϕ_j , given by a Gaussian distribution centered at ϕ_j , as shown in Eq. 2. If two feature vectors are not close enough to each other, the probability will be infinitesimal. Temperature coefficient T controls the number of neighbors being considered.

After the computation of similarity s_{ij} , binary cross-entropy loss is calculated accordingly to enforce the network giving similar predictions for image data with similar feature vectors, as shown in Eq. 4 and Eq. 5.

$$\mathcal{L}_{BCE} = -\frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m \{A_{ij} \log(\hat{\mathbf{y}}_i^\top \cdot \hat{\mathbf{y}}_j) + (1 - A_{ij}) \log(1 - \hat{\mathbf{y}}_i^\top \cdot \hat{\mathbf{y}}_j)\} \quad (4)$$

$$A_{ij} = \begin{cases} 1 & \text{if } s_{ij} > \tau \\ s_{ij} & \text{otherwise} \end{cases} \quad (5)$$

where $\hat{\mathbf{y}}_i = \text{softmax}(f_{c1}(\phi_i)) \in R^{c_u}$, the output from linear layer f_{c1} for feature vector ϕ_i followed by a *softmax* function. In our proposed algorithm, we wish
170 to update the network once per batch. Therefore, $\mathcal{L}_{BCE}$ is calculated for image data in a batch and m is the number of unlabeled data in a batch. Similarly,

the calculation of s_{ij} is done for each batch, instead of the whole dataset. A_{ij} equals to 1 if two image data x_i and x_j are considered to be similar, i.e., s_{ij} is larger than the threshold value τ . Here the inner product between $\hat{\mathbf{y}}_i$ and $\hat{\mathbf{y}}_j$ measures the similarity between network predictions for image data x_i and x_j .

3.2.2. Mixed Label Classification

Besides pairwise similarity learning for unlabeled data, discriminative feature learning for classification is essential as label information is known for X_l . We first calculate cross-entropy loss $\mathcal{L}_l$ between the network predictions and Y_l for all data in X_l . For cross-entropy loss $\mathcal{L}_u$ on unlabeled data, predictions $f_{c1}(\phi_i)$ produced by the network in the previous pairwise similarity learning can be used to infer pseudo labels. Therefore, cross-entropy loss is computed between the network predictions and true labels for X_l together with pseudo labels for X_u . One more linear layer f_{c2} is added to the network to map feature vectors to label space of dimension $c_l + c_u$. The computation of $\mathcal{L}_{CE}$ is shown in Eq. 6.

$$\begin{aligned}\mathcal{L}_{CE} &= \mathcal{L}_l + \mathcal{L}_u \\ &= -\frac{1}{n} \left(\sum_{i: x_i \in X_l} \log(\hat{\mathbf{y}}_i)_{y_i} + \lambda_{pl} \sum_{i: x_i \in X_u} \log(\hat{\mathbf{y}}_i)_{p_i} \right)\end{aligned}\quad (6)$$

where n is the batch size, i.e., the total number of labeled data and unlabeled data in a batch, $\hat{\mathbf{y}}_i = \text{softmax}(f_{c2}(\phi_i)) \in R^{c_l+c_u}$, the output from linear layer f_{c2} for feature vector ϕ_i followed by a *softmax* function, y_i is the ground-truth label, λ_{pl} is the weight coefficient for pseudo labels and the pseudo label p_i for each unlabeled data is inferred with Eq. 7.

$$p_i = c_l + \arg \max \{ \text{softmax}(f_{c1}(\phi_i)) \} \quad (7)$$

Pseudo label p_i is inferred by the output from the first linear layer f_{c1} incremented by c_l , as pseudo labels are mixed together with ground-truth labels to calculate the loss. Similar to [14], consistency loss is added on the output from linear layer f_{c1} and f_{c2} to enforce the stability and robustness of our algorithm,

as shown in Eq. 8.

$$\begin{aligned}\mathcal{L}_{MSE} = & \frac{1}{n-m} \sum_{i:x_i \in X_l} \|\text{softmax}(f_{c2}(\phi_i)) - \text{softmax}(f_{c2}(\hat{\phi}_i))\|^2 \\ & + \frac{1}{m} \sum_{i:x_i \in X_u} \|\text{softmax}(f_{c1}(\phi_i)) - \text{softmax}(f_{c1}(\hat{\phi}_i))\|^2\end{aligned}\quad (8)$$

where $\hat{\phi}_i$ denotes the feature vector of $\hat{x}_i$, a random-augmented counterpart of x_i .

3.2.3. Self-Supervised Learning

Pairwise similarity learning builds the pairwise relationship between unlabeled data based on the learned feature space, which is primarily supervised by the label information of X_l , shown in Eq. 6. However, the extracted features solely depending on image data from labeled classes are highly likely to be biased and its generalization capability for unlabeled data will be degraded. To alleviate such problem, we propose to given additional supervision signal other than the label information, which is only available for limited number of data. In [16], the network is trained to predict rotational geometric transformation that has been applied to the image data, serving the role of self-supervision. We incorporate the rotation-based self-supervision signal into our algorithm by requiring the network to classify image categories and rotational transformations that has been applied simultaneously.

Every image $x_i \in X$ is rotated by $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$, and the rotated images $\mathbb{X}$ are fed into the same network to get feature vectors, as shown in Fig. 1. The third linear layer f_{c3} is added to map feature vectors of rotated images to surrogate label space of dimension 4. Rotation loss $\mathcal{L}_{Rot}$ is calculated by computing the cross-entropy loss between network prediction $\hat{\mathbf{y}}_i$ and rotational surrogate label $r = \{0, 1, 2, 3\}$ for images after rotational transformation of $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$ respectively, as shown in Eq. 9.

$$\mathcal{L}_{Rot} = -\frac{1}{n} \sum_{i:x_i \in \mathbb{X}} \log(\hat{\mathbf{y}}_i)_{r_i} \quad (9)$$

where $\hat{\mathbf{y}}_i = \text{softmax}(f_{c3}(\phi_i)) \in R^4$, the output from linear layer f_{c3} for feature vector ϕ_i of all rotated images followed by a *softmax* function.

The overall objective function is computed in Eq. 10.

$$\mathcal{L}_{all} = (1 - \lambda_{Rot})(\mathcal{L}_{BCE} + \mathcal{L}_{CE} + \beta\mathcal{L}_{MSE}) + \lambda_{Rot}\mathcal{L}_{Rot} \quad (10)$$

where λ_{Rot} controls the relative importance between auxiliary self-supervision and explicit label information, and β determines the weight of consistency loss. The form of $\mathcal{L}_{MSE}$ determines that the magnitude of the loss will be minuscule due to the power operation on probability distribution, which is different from the other three loss terms. Moreover, ideal model is expected to give same predictions for input data x_i and its random-augmented counterpart $\hat{x}_i$ such that $\mathcal{L}_{MSE}$ should be close to zero. Therefore, β is normally set to a larger value to balance the overall loss. The pseudocode for the proposed algorithm is summarized in Algorithm 1.

4. Experiments

4.1. Datasets

- **CIFAR-10** dataset [51] is one commonly-used benchmark dataset for visual classification. There are total 60,000 color images of size 32×32 , evenly distributed among ten different classes. For each class, 5000 images belong to training data and the rest 1000 images belong to testing data.
- **CIFAR-100** dataset [51] is similar to CIFAR-10 in terms of dataset size, but CIFAR-100 is much more challenging than CIFAR-10 as there are more diverse images in CIFAR-100. There are total 60,000 colorful images of size 32×32 , evenly distributed among 100 different classes. For each class, 500 images belong to training data and the rest 100 images belong to testing data.
- **SVHN** (Street View House Numbers) dataset [52] is a real-world digit number recognition dataset, comprising 73,257 and 26,032 colorful images of size 32×32 for training and testing data respectively. Pure digit

Algorithm 1 End-to-end Novel Visual Categories Learning via Auxiliary Self-Supervision

```

1: Input: Labeled training images from known classes and the corresponding
   labels  $\{X_l, Y_l\}$ , unlabeled training images  $X_u$  from novel classes, the number
   of all possible labeled classes  $c_l$ , the number of all possible novel classes  $c_u$ ,
   temperature coefficient  $T$ , threshold value  $\tau$ , weight coefficient  $\lambda_{pl}$ ,  $\lambda_{Rot}$ 
   and  $\beta$ , convolutional layers of random initialized neural network  $g_\theta(\cdot)$ , three
   linear layers  $f_{c1}(\cdot), f_{c2}(\cdot), f_{c3}(\cdot)$ 
2: for  $epoch \leftarrow 1$  to # of training epochs do
3:   for sampled minibatch  $B$  do
4:      $\phi_{i \in B} \leftarrow g_\theta(x_i)$ 
5:     for  $i, j \in B \cap X_u$  do
6:        $p_{i|j} \leftarrow \frac{\exp(-\|\phi_j - \phi_i\|^2/T^2)}{\sum_{k \neq j} \exp(-\|\phi_j - \phi_k\|^2/T^2)}$ 
7:        $s_{ij} \leftarrow \frac{1}{2}(p_{i|j} + p_{j|i})$ 
8:        $\mathcal{L}_{BCE} \leftarrow \sum L(A_{ij}, f_{c1}(\phi_i), f_{c1}(\phi_j))$ , where  $A_{ij}$  is defined in 5
9:        $p_i \leftarrow c_l + \arg \max\{softmax(f_{c1}(\phi_i))\}$ 
10:       $\mathcal{L}_{CE} \leftarrow \sum_{i \in B \cap X_l} L(y_i, f_{c2}(\phi_i)) + \lambda_{pl} \sum_{i \in B \cap X_u} L(p_i, f_{c2}(\phi_i))$ 
11:       $\hat{x}_{i \in B} \leftarrow$  randomly augment  $x_i$ 
12:       $\hat{\phi}_i \leftarrow g_\theta(\hat{x}_i)$ 
13:       $\mathcal{L}_{MSE} \leftarrow \sum_{i \in B \cap X_l} L(f_{c2}(\phi_i), f_{c2}(\hat{\phi}_i)) + \sum_{i \in B \cap X_u} L(f_{c1}(\phi_i), f_{c1}(\hat{\phi}_i))$ 
14:       $x_{i \in B}^r \leftarrow$  rotate  $x_i$  by  $0^\circ, 90^\circ, 180^\circ, 270^\circ$ 
15:       $\phi_i^r \leftarrow g_\theta(x_i^r)$ 
16:       $r_{i \in B} \leftarrow \{0, 1, 2, 3\}$ 
17:       $\mathcal{L}_{Rot} \leftarrow \sum_{i \in B} L(r_i, f_{c3}(\phi_i^r))$ 
18:       $loss \leftarrow (1 - \lambda_{Rot})(\mathcal{L}_{BCE} + \mathcal{L}_{CE} + \beta \mathcal{L}_{MSE}) + \lambda_{Rot} \mathcal{L}_{Rot}$ 
19:      update  $f_{c1}(\cdot), f_{c2}(\cdot), f_{c3}(\cdot)$  and  $g_\theta(\cdot)$   $\triangleright$  back-propagation
20: Output: Model parameters  $\theta$  of  $g_\theta(\cdot)$  and  $f_{c1}(\cdot), f_{c2}(\cdot), f_{c3}(\cdot)$ 

```

205 datasets, e.g., MNIST, are not suitable for the rotation-based auxiliary
task in the proposed algorithm as pairs like “6” and “9” will be confusing
and introduce noise to the model. However, the digits in SVHN dataset
are not “clean” as the data are cropped from the original house-number
images. Therefore, there are “distracting” digits surrounding the digit of
210 interest. Those extra “distracting” digits improve the distinction between
confusing pairs under various rotational transformations.

4.2. Implementation

The network architecture used is ResNet-18 [53]. For CIFAR-10 and SVHN
dataset, half of the training data is used as labeled data and the other half is
used as unlabeled data, i.e., $c_l = c_u = 5$. For CIFAR-100, 80 classes of training
images are used as labeled data and the rest 20 classes of training images are
utilized as unlabeled images from novel categories, i.e., $c_l = 80$ and $c_u = 20$. For
the hyper-parameter setting, λ_{Rot} is set to be 0.8 in all datasets to enforce the
regularization effects and we follow the settings reported in [14] for the weight
coefficient λ_{pl} of pseudo labels and the weight β of consistency loss. λ_{pl} and β
take the form of a ramp-up function $w(t)$ along the training process, computed
with Eq. 11.

$$w(t) = \exp(-5(1 - \frac{t}{\mathbb{T}})^2) \quad (11)$$

$$\lambda_{pl} = A_{pl} \cdot w(t), \quad \beta(t) = A_{mse} \cdot w(t)$$

where t is the current training epoch and $\mathbb{T}$ is the total number of epochs in one
period. $\mathbb{T}$ is set to be $\{50, 300, 80\}$, A_{mse} is set to be $\{5, 25, 50\}$ and $A_{pl} = 0.05$
215 for $\{\text{CIFAR-10, CIFAR-100, SVHN}\}$. The total number of training epochs is
set to be $\{200, 400, 200\}$ and initial learning rate is set to be $\{0.05, 0.1, 0.05\}$ for
 $\{\text{CIFAR-10, CIFAR-100, SVHN}\}$. For all datasets, we use Stochastic Gradient
Descent (SGD) optimizer for back-propagation and the learning rate follows
a cosine annealing scheme [54]. The batch size is set to be $\{128, 256, 128\}$,
220 temperature coefficient T in Eq. 4 is set to be $\{1, 0.5, 1\}$, and threshold value τ
in Eq. 5 is set to be 0.01 for $\{\text{CIFAR-10, CIFAR-100, SVHN}\}$. A smaller value

set for temperature coefficient T of CIFAR-100 is due to the fact that there are less similar pairs of data per batch for CIFAR-100 dataset than those for the other two datasets. For threshold value τ , it determines the relative importance between the two terms of $\mathcal{L}_{BCE}$ in Eq. 4. If the threshold value τ is high, it means that the second term $(1 - A_{ij}) \log(1 - \hat{\mathbf{y}}_i^\top \cdot \hat{\mathbf{y}}_j)$ will dominate $\mathcal{L}_{BCE}$, which will encourage the model to give dissimilar predictions for most pairs. On the contrary, a low threshold value τ will encourage the model to give similar predictions for most pairs as the first term $A_{ij} \log(\hat{\mathbf{y}}_i^\top \cdot \hat{\mathbf{y}}_j)$ will dominate the loss. For datasets used in the experiments, similar pairs are much less than dissimilar pairs as none of the datasets is for binary classification. Therefore, we set τ to a pretty small value to alleviate such imbalance such that the network can give more similar predictions. The implementation details are given in Table 1.

Table 1: Implementation details of the experiments

	CIFAR-10	CIFAR-100	SVHN
Initial Learning Rate	0.05	0.1	0.05
Batch Size	128	256	128
Total Training Epochs	200	400	200
c_l	5	80	5
c_u	5	20	5
$\mathbb{T}$	50	300	80
A_{mse}	5	25	50
T	1	0.5	1
Network	ResNet-18		
Optimizer	SGD		
LR Annealing Scheme	cosine		
λ_{Rot}	0.8		
A_{pl}	0.05		
τ	0.01		

4.3. Evaluation

As the objective is to cluster X_u , our algorithm is evaluated with clustering accuracy ACC metric in Eq. 12:

$$ACC = \max_{perm \in P} \frac{1}{N-l} \sum \delta_{\hat{y}_i, perm(y_i)} \quad (12)$$

235 where $\hat{y}_i$ is the network prediction for unlabeled data $x_i \in X_u$, i.e., $\hat{y}_i = \arg \max \{softmax(f_{c1}(\phi_i))\}$ and y_i is the ground-truth label. P denotes all possible permutations for the indices of c_u clusters. δ denotes Kronecker delta function.

4.4. Experiment Results

240 4.4.1. Clustering accuracy

The clustering capability of our proposed algorithm is compared with previous state-of-the-art methods in Table 2. Six methods listed in the bottom part of Table 2 including our proposed algorithm have utilized self-supervised learning, whereas those listed in the upper part haven't. Methods in the bottom
245 part of Table 2 generally perform better than those in the upper part, which indicates that it is beneficial to use self-supervised learning for the clustering task of concern. Among all methods that have utilized self-supervised learning, our proposed algorithm achieves the best results on all three datasets. Our results surpass the latest best results of AutoNovel [14] by 1.9%, 6.1% and 0.2% on all
250 three datasets.

4.4.2. Mixed classification performance

Besides the clustering capability of the trained model on unlabeled training images, classifying new and unseen images from either "old" classes or "novel" classes is of great significance for practical applications. Therefore, we further
255 evaluate our proposed algorithm together with other state-of-the-art ones on classification performance on testing images in these three datasets. In this case, the output from linear layer f_{c2} is evaluated, and the results are listed in Table 3 and Table 4. On CIFAR datasets in Table 3, our proposed algorithm achieves

Table 2: Clustering Accuracy on unlabeled training data (%). w/ SSL denotes self-supervised learning is used in the pre-training stage of the network. All reported results are using the same network architecture and averaged over ten runs.

Dataset	CIFAR-10	CIFAR-100	SVHN
k -means [55]	65.5 ± 0.0	56.6 ± 1.6	42.6 ± 0.0
KCL[12]	66.5 ± 3.9	14.3 ± 1.3	21.4 ± 0.6
MCL[13]	64.2 ± 0.1	21.3 ± 3.4	38.6 ± 10.8
DTC[33]	87.5 ± 0.3	56.7 ± 1.2	60.9 ± 1.6
k -means [55] w/ SSL	72.5 ± 0.0	56.3 ± 1.7	46.7 ± 0.0
KCL[12] w/ SSL	72.3 ± 0.2	42.1 ± 1.8	65.6 ± 4.9
MCL[13] w/ SSL	70.9 ± 0.1	21.5 ± 2.3	53.1 ± 0.3
DTC[33] w/ SSL	88.7 ± 0.3	67.3 ± 1.2	75.7 ± 0.4
AutoNovel [14]	91.7 ± 0.7	75.2 ± 4.2	95.2 ± 0.2
Ours	93.6 ± 0.6	81.3 ± 1.9	95.4 ± 0.3

the best results for all categories. Our algorithm has outperformed others on both “old” and “novel” classes. It indicates that our algorithm is not biased to either category, as the trained network generalizes equally well on both. For SVHN dataset in Table 4, our algorithm achieves the best result on “old” classes, but AutoNovel [14] beats our algorithm on “novel” classes. As classifying SVHN dataset is a relatively easier task compared with CIFAR datasets, we deduce the reason is that overfitting has degraded the performance. In AutoNovel [14], label information is only used to tune the last block of network while most parameters are frozen after self-supervised learning, which limits the learning capability but meanwhile largely avoids overfitting. Comparing with AutoNovel [14], our algorithm is able to capture more information as the whole network is updating during the training process, whereas overfitting problem may arise when dealing with relatively easy datasets. Despite its suboptimal performance on “novel” classes, our algorithm still achieves the best result on the whole

testing dataset including all “old” classes and “novel” classes.

Table 3: Mixed Classification Performance on CIFAR datasets (%). We evaluate the trained network on testing data for each dataset. “Old” refers to testing images from classes that have been trained by training images with label information, while “Novel” refers to testing images from classes that no label information is used during training. “All” refers to all testing images in the dataset.

Dataset	CIFAR-10			CIFAR-100		
Category	Old	Novel	All	Old	Novel	All
KCL[12] w/ SSL	79.3 ± 0.6	60.1 ± 0.6	69.8 ± 0.1	23.4 ± 0.3	29.4 ± 0.3	24.6 ± 0.2
MCL[13] w/ SSL	81.4 ± 0.4	64.8 ± 0.4	73.1 ± 0.1	18.2 ± 0.3	18.0 ± 0.1	18.2 ± 0.2
DTC[33] w/ SSL	58.7 ± 0.6	78.6 ± 0.2	68.7 ± 0.3	47.6 ± 0.2	49.1 ± 0.2	47.9 ± 0.2
AutoNovel [14]	90.6 ± 0.2	88.8 ± 0.2	89.7 ± 0.1	71.2 ± 0.1	56.8 ± 0.3	68.3 ± 0.1
Ours	94.6 ± 0.4	90.4 ± 0.4	92.5 ± 0.2	76.0 ± 0.2	66.7 ± 1.8	74.1 ± 0.5

Table 4: Mixed Classification Performance on SVHN dataset (%)

Dataset	SVHN		
Category	Old	Novel	All
KCL[12] w/ SSL	90.3 ± 0.3	65.0 ± 0.5	81.0 ± 0.1
MCL[13] w/ SSL	94.0 ± 0.2	48.6 ± 0.3	77.2 ± 0.1
DTC[33] w/ SSL	90.5 ± 0.3	72.8 ± 0.2	84.0 ± 0.1
AutoNovel [14]	96.3 ± 0.1	96.1 ± 0.0	96.2 ± 0.1
Ours	97.6 ± 0.2	94.8 ± 0.3	96.5 ± 0.1

4.4.3. Ablation Study

275 To evaluate each module of our proposed algorithm, ablation study is performed on CIFAR-100 dataset and the results are listed in Table 5.

Self-supervised learning: First of all, the bias between mixed classification performance on “old” classes and “novel” classes are largely alleviated due to

Table 5: Ablation Study (%)

	Ablation	Clustering	Mixed Classification		
		ACC	Old	Novel	All
	Ours	80.99	75.90	67.00	74.09
Variant (1)	Ours without $\mathcal{L}_{Rot}$	70.98	76.81	57.85	72.92
Variant (2)	Ours without $\mathcal{L}_{MSE}$	79.97	75.09	65.45	73.16
Variant (3)	Ours without $\mathcal{L}_l$	51.00	4.11	39.45	8.34
Variant (4)	Ours without $\mathcal{L}_u$	73.49	77.75	47.90	62.20
Variant (5)	Ours without $\mathcal{L}_{BCE}$	33.51	72.01	31.40	63.52
Variant (6a)	Ours with cosine similarity [†]	75.56	75.39	63.45	72.96
Variant (6b)	Ours with rank statistics [★]	74.14	75.59	61.75	72.79
Variant (7)	Ours with SSL as initialization [*]	74.95	71.29	58.75	68.75

[†] We replace proposed conditional probability with cosine similarity, i.e., $s_{ij} = \frac{\phi_i \cdot \phi_j}{||\phi_i|| \times ||\phi_j||}$, as similarity measure for the calculation of $\mathcal{L}_{BCE}$.

[★] The proposed conditional probability is replaced with rank statistics in [14].

^{*} We follow the pre-training and fine-tuning schemes in [14], i.e., self-supervised learning is only utilized independently for the network initialization.

the use of self-supervision, as the discrepancy, i.e., the classification accuracy
gap between “old” classes and “novel” classes, 8.90% of ours and 12.54% of
variant (7) are lower than 18.96% of variant (1). Secondly, it is observed that
self-supervised learning can always boost the performance on both clustering
and mixed classification for “novel” classes by comparing variant (7) and ours
with variant (1). Lastly, the results have verified our argument that leveraging
self-supervision simultaneously with pairwise similarity information can further
boost model’s learning capability as our algorithm outperforms variant (7) by
a significant margin on all evaluation metrics. The difference between ours and
variant (7) is the way how self-supervision is utilized: self-supervision and pair-

wise similarity information are jointly training the network in our algorithm
 290 while self-supervision is only used for pre-training in variant (7).

Pairwise similarity learning: Firstly, it is observed that pairwise similarity learning is the most crucial module for novel categories learning as only 33.51% and 31.40% accuracies have been achieved by variant (5) for clustering and mixed classification on “novel” classes respectively, which are the worst
 295 results among all variants. Secondly, the use of conditional probability as pairwise similarity measurement can capture such pairwise relationship much more accurately in the feature space than commonly-used cosine similarity. It can be observed by comparing ours with variant (6a) that the our algorithm has outperformed variant (6a) by significant margins of 5.43% and 3.55% for clustering and mixed classification on “novel” classes respectively. Moreover, the
 300 rank statistics proposed in [14] are also compared with ours, as shown in variant (6b). The result indicates that the rank statistics are less robust to noises in the feature space trained with the proposed learning framework, as the clustering and mixed classification accuracy of variant (6b) are less than 75% and 62% on “novel” classes respectively, which are even poorer than those of variant (6a).
 305

Mixed Classification: By comparing variant (3) and (4) with ours, we notice that mixed classification plays an important role for both “old” and “novel” classes learning. Without $\mathcal{L}_u$, the pseudo labels inferred by output from f_{c1} are not utilized. Both clustering accuracy and mixed classification performance on
 310 “novel” classes have dropped and the trained network is largely biased to “old” classes during mixed classification. Without $\mathcal{L}_l$, i.e., the ground-truth label information for “old” classes is removed, all evaluation metrics have dropped a lot. The effectiveness of consistency loss $\mathcal{L}_{MSE}$ has been demonstrated by comparing ours with variant (2).

315 In Fig. 2, 2D t-SNE visualizations for feature vectors ϕ of training images from novel classes are demonstrated for models with different training schemes. Without $\mathcal{L}_{Rot}$, i.e., only pairwise similarity information and ground-truth label information are supervising the model, feature vectors belonging to different classes are not well separated and some parts are connected with each other

Figure 2: 2D t-SNE visualizations of embeddings ϕ for CIFAR-100 unlabeled training data under different training schemes. Each single point represents one unlabeled training image and different colors correspond to ground-truth labels. Recommend to view in color version.

as shown in Fig. 2a. For the case when pairwise similarity is measured by
 cosine similarity, there are more feature vectors clustered to the wrong groups
 compared with our proposed algorithm as shown in Fig. 2b, indicating using
 conditional probability as pairwise similarity criterion is more accurate than
 cosine similarity.

5. Conclusion

In this paper, we have proposed a new end-to-end novel visual categories
 learning algorithm by utilizing self-supervision as an auxiliary task. Two major

issues inherent in prior relevant methods reported in literature are limitation
 of learning capability due to the difficulty for end-to-end training and pairwise
 330 similarity pseudo labels' vulnerability to noise and bias. To address these two
 problems, we propose to use surrogate labels from self-supervision together with
 pairwise similarity pseudo labels to train the model simultaneously, which pro-
 vides extra supervision and regularization for end-to-end learning. Moreover,
 we propose to construct pairwise similarity pseudo labels by exploiting robust
 335 local data structure information in noisy feature space. As such, the quality of
 pairwise pseudo labels can be further improved. The ablation study has demon-
 strated the contribution from each single module and has shown the efficacy
 of proposed designs. Experiments have been conducted on three commonly-
 used visual datasets i.e., CIFAR-10, CIFAR-100 and SVHN. The results have
 340 indicated the effectiveness of our algorithm as the proposed algorithm has out-
 performed previous state-of-the-art algorithms significantly.

The motivation behind this work is to alleviate the bias arising from lack of
 annotations for images from novel categories. According to [14], self-supervised
 learning does not demonstrate its effectiveness on ImageNet and OmniGlot [56]
 345 for novel categories learning. The reason behind is that the diversity and abun-
 dance of those large datasets have already reduced the bias between different
 categories. Therefore, the focus of novel categories learning problem for such
 kind of datasets may not be on eliminating the bias between various categories.
 To further explore effective algorithms of novel visual categories learning for
 350 abundant and diverse datasets is one future research direction. Moreover, the
 novel categories learning problem discussed in this paper assumes that the num-
 ber of novel categories is known, which may not be satisfied for realistic applica-
 tions. In a more challenging setting, the number of novel categories/number of
 clusters for unlabeled data is unknown and needs to be approximated. How to
 355 determine the number of clusters for novel categories learning is an important
 question worth further research on.

References

- [1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, Li Fei-Fei, ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.
- [2] O. Chapelle, B. Schölkopf, A. Zien (Eds.), Semi-Supervised Learning, Adaptive Computation and Machine Learning, MIT Press, Cambridge, Mass, 2006.
- [3] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A Simple Framework for Contrastive Learning of Visual Representations, arXiv:2002.05709 [cs, stat] (Jun. 2020). arXiv:2002.05709.
- [4] X. Zhai, A. Oliver, A. Kolesnikov, L. Beyer, S4L: Self-Supervised Semi-Supervised Learning, arXiv:1905.03670 [cs] (Jul. 2019). arXiv:1905.03670, doi:10.1109/iccv.2019.00156.
- [5] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, Q. V. Le, Unsupervised Data Augmentation for Consistency Training, arXiv:1904.12848 [cs, stat] (Sep. 2019). arXiv:1904.12848.
- [6] A. Iscen, G. Tolias, Y. Avrithis, O. Chum, Label Propagation for Deep Semi-Supervised Learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5070–5079. doi:10.1109/cvpr.2019.00521.
- [7] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, arXiv:1703.01780 [cs, stat] (Apr. 2018). arXiv:1703.01780.
- [8] T. Miyato, S.-i. Maeda, M. Koyama, S. Ishii, Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning, arXiv:1704.03976 [cs, stat] (Jun. 2018). arXiv:1704.03976, doi:10.1109/tpami.2018.2858821.

- 385 [9] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, C. Raffel, MixMatch: A Holistic Approach to Semi-Supervised Learning, arXiv:1905.02249 [cs, stat] (Oct. 2019). **arXiv:1905.02249**.
- [10] S. Laine, T. Aila, Temporal Ensembling for Semi-Supervised Learning, arXiv:1610.02242 [cs] (Mar. 2017). **arXiv:1610.02242**.
- 390 [11] D.-H. Lee, Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks 6.
- [12] Y.-C. Hsu, Z. Lv, Z. Kira, Learning to cluster in order to transfer across domains and tasks, in: International Conference on Learning Representations (ICLR), 2018.
- 395 URL <https://openreview.net/forum?id=ByRWCqvT->
- [13] J. S. P. O. Z. K. Yen-Chang Hsu, Zhaoyang Lv, Multi-class classification without multi-class labels, in: International Conference on Learning Representations (ICLR), 2019.
- URL <https://openreview.net/forum?id=SJzR2iRcK7>
- 400 [14] K. Han, S.-A. Rebuffi, S. Ehrhardt, A. Vedaldi, A. Zisserman, Automatically Discovering and Learning New Visual Categories with Ranking Statistics, in: International Conference on Learning Representations, 2019.
- [15] L. Jing, Y. Tian, Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey, arXiv:1902.06162 [cs] (Feb. 2019). **arXiv:1902.06162**, doi:10.1109/tpami.2020.2992393.
- 405
- [16] S. Gidaris, P. Singh, N. Komodakis, Unsupervised Representation Learning by Predicting Image Rotations, arXiv:1803.07728 [cs] (Mar. 2018). **arXiv:1803.07728**.
- [17] G. Larsson, M. Maire, G. Shakhnarovich, Colorization as a Proxy Task for Visual Understanding, arXiv:1703.04044 [cs] (Aug. 2017). **arXiv:1703.04044**, doi:10.1109/cvpr.2017.96.
- 410

- [18] M. Noroozi, P. Favaro, Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles, arXiv:1603.09246 [cs] (Aug. 2017). **arXiv:1603.09246**, doi:10.1007/978-3-319-46466-4_5.
- 415 [19] S.-A. Rebuffi, S. Ehrhardt, K. Han, A. Vedaldi, A. Zisserman, Semi-Supervised Learning with Scarce Annotations, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 3294–3302. doi:10.1109/CVPRW50498.2020.00389.
- 420 [20] J. Cook, I. Sutskever, A. Mnih, G. Hinton, Visualizing Similarity Data with a Mixture of Maps 8.
- [21] V. Verma, A. Lamb, J. Kannala, Y. Bengio, D. Lopez-Paz, Interpolation Consistency Training for Semi-Supervised Learning, arXiv:1903.03825 [cs, stat] (May 2019). **arXiv:1903.03825**, doi:10.24963/ijcai.2019/504.
- 425 [22] S. Park, J.-K. Park, S.-J. Shin, I.-C. Moon, Adversarial Dropout for Supervised and Semi-supervised Learning, arXiv:1707.03631 [cs] (Sep. 2017). **arXiv:1707.03631**.
- [23] B. Athiwaratkun, M. Finzi, P. Izmailov, A. G. Wilson, There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average, arXiv:1806.05594 [cs, stat] (Feb. 2019). **arXiv:1806.05594**.
- 430 [24] W. Shi, Y. Gong, C. Ding, Z. Ma, X. Tao, N. Zheng, Transductive Semi-Supervised Deep Learning Using Min-Max Features, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision – ECCV 2018, Vol. 11209, Springer International Publishing, Cham, 2018, pp. 311–327. doi:10.1007/978-3-030-01228-1_19.
- 435 [25] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, The Regents of the University of California, 1967.

- [26] J. Shi, J. Malik, Normalized cuts and image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (8) (2000) 888–905. doi:10.1109/34.868688.
- [27] D. Comaniciu, P. Meer, Mean shift: A robust approach toward feature space analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (5) (2002) 603–619. doi:10.1109/34.1000236.
- [28] A. Y. Ng, M. I. Jordan, Y. Weiss, On Spectral Clustering: Analysis and an algorithm, in: T. G. Dietterich, S. Becker, Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14*, MIT Press, 2002, pp. 849–856.
- [29] J. Xie, R. Girshick, A. Farhadi, Unsupervised Deep Embedding for Clustering Analysis, in: *International Conference on Machine Learning*, PMLR, 2016, pp. 478–487.
- [30] J. Chang, L. Wang, G. Meng, S. Xiang, C. Pan, Deep Adaptive Image Clustering, in: *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Venice, 2017, pp. 5880–5888. doi:10.1109/ICCV.2017.626.
- [31] M. Caron, P. Bojanowski, A. Joulin, M. Douze, Deep Clustering for Unsupervised Learning of Visual Features, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2018, pp. 139–156. doi:10.1007/978-3-030-01264-9_9.
- [32] N. Mrabah, N. M. Khan, R. Ksantini, Z. Lachiri, Deep clustering with a Dynamic Autoencoder: From reconstruction towards centroids construction, *Neural Networks* 130 (2020) 206–228. doi:10.1016/j.neunet.2020.07.005.
- [33] K. Han, A. Vedaldi, A. Zisserman, Learning to Discover Novel Visual Categories via Deep Transfer Clustering, in: *Proceedings of the IEEE/CVF*

International Conference on Computer Vision, 2019, pp. 8401–8409. doi:
10.1109/iccv.2019.00849.

- 470 [34] M. Śmieja, L. Struski, M. A. T. Figueiredo, A classification-based approach
to semi-supervised clustering with pairwise constraints, *Neural Networks*
127 (2020) 193–203. doi:10.1016/j.neunet.2020.04.017.
- [35] S.-A. Rebuffi, S. Ehrhardt, K. Han, A. Vedaldi, A. Zisserman, LSD-
C: Linearly Separable Deep Clusters, arXiv:2006.10039 [cs] (Jun. 2020).
arXiv:2006.10039.
- 475 [36] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked
Denoising Autoencoders: Learning Useful Representations in a Deep Net-
work with a Local Denoising Criterion 38.
- [37] H. Larochelle, D. Erhan, Y. Bengio, Zero-data learning of new tasks, in:
Proceedings of the 23rd national conference on Artificial intelligence - Vol-
480 ume 2, AAAI’08, AAAI Press, Chicago, Illinois, 2008, pp. 646–651.
- [38] C. H. Lampert, H. Nickisch, S. Harmeling, Attribute-Based Classification
for Zero-Shot Visual Object Categorization, *IEEE Transactions on Pattern
Analysis and Machine Intelligence* 36 (3) (2014) 453–465, conference Name:
IEEE Transactions on Pattern Analysis and Machine Intelligence. doi:
485 10.1109/TPAMI.2013.140.
- [39] M. Rohrbach, M. Stark, B. Schiele, Evaluating knowledge transfer and
zero-shot learning in a large-scale setting, in: *CVPR 2011*, IEEE, Col-
orado Springs, CO, USA, 2011, pp. 1641–1648. doi:10.1109/CVPR.2011.
5995627.
490 URL <http://ieeexplore.ieee.org/document/5995627/>
- [40] X. Yu, Y. Aloimonos, Attribute-Based Transfer Learning for Object Cat-
egorization with Zero/One Training Example, in: K. Daniilidis, P. Mara-
gos, N. Paragios (Eds.), *Computer Vision – ECCV 2010*, Lecture Notes

- in Computer Science, Springer, Berlin, Heidelberg, 2010, pp. 127–140.
 495 doi:10.1007/978-3-642-15555-0_10.
- [41] X. Xu, F. Shen, Y. Yang, D. Zhang, H. T. Shen, J. Song, Matrix Tri-Factorization with Manifold Regularizations for Zero-Shot Learning, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2007–2016, iSSN: 1063-6919. doi:10.1109/CVPR.2017.
 500 217.
- [42] Z. Ding, M. Shao, Y. Fu, Low-Rank Embedded Ensemble Semantic Dictionary for Zero-Shot Learning, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, 2017, pp. 6005–6013. doi:10.1109/CVPR.2017.636.
 505 URL <http://ieeexplore.ieee.org/document/8100119/>
- [43] Y. Xian, C. H. Lampert, B. Schiele, Z. Akata, Zero-Shot Learning – A Comprehensive Evaluation of the Good, the Bad and the Ugly, arXiv:1707.00600 [cs]ArXiv: 1707.00600 (Sep. 2020).
 URL <http://arxiv.org/abs/1707.00600>
- 510 [44] C. H. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 951–958, iSSN: 1063-6919. doi:10.1109/CVPR.2009.5206594.
- [45] B. Romera-Paredes, P. H. S. Torr, An Embarrassingly Simple Approach
 515 to Zero-Shot Learning, in: R. S. Feris, C. Lampert, D. Parikh (Eds.), Visual Attributes, Springer International Publishing, Cham, 2017, pp. 11–30, series Title: Advances in Computer Vision and Pattern Recognition. doi:10.1007/978-3-319-50077-5_2.
 URL http://link.springer.com/10.1007/978-3-319-50077-5_2
- 520 [46] Y. Atzmon, G. Chechik, Probabilistic AND-OR Attribute Grouping for Zero-Shot Learning, arXiv:1806.02664 [cs]ArXiv: 1806.02664 (Jul. 2018).
 URL <http://arxiv.org/abs/1806.02664>

- [47] Z. Akata, S. Reed, D. Walter, Honglak Lee, B. Schiele, Evaluation of output embeddings for fine-grained image classification, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Boston, MA, USA, 2015, pp. 2927–2936. doi:10.1109/CVPR.2015.7298911.
URL <http://ieeexplore.ieee.org/document/7298911/>
- [48] R. Socher, M. Ganjoo, C. D. Manning, A. Ng, Zero-Shot Learning Through Cross-Modal Transfer, 2013, pp. 935–943.
URL <http://papers.neurips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer>
- [49] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, J. Dean, Zero-shot learning by convex combination of semantic embeddings, 2014.
URL <https://collaborate.princeton.edu/en/publications/zero-shot-learning-by-convex-combination-of-semantic-embeddings>
- [50] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS’13, Curran Associates Inc., Lake Tahoe, Nevada, 2013, pp. 3111–3119.
- [51] A. Krizhevsky, Learning multiple layers of features from tiny images, Tech. rep. (2009).
- [52] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng, Reading Digits in Natural Images with Unsupervised Feature Learning, in: NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011, 2011.
- [53] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

- 550 [54] I. Loshchilov, F. Hutter, SGDR: Stochastic Gradient Descent with Warm
Restarts, arXiv:1608.03983 [cs, math] (May 2017). `arXiv:1608.03983`.
- [55] L. M. L. Cam, J. Neyman, Proceedings of the Fifth Berkeley Symposium on
Mathematical Statistics and Probability: Weather Modification, University
of California Press, 1967.
- 555 [56] B. M. Lake, R. Salakhutdinov, J. B. Tenenbaum, Human-level concept
learning through probabilistic program induction, Science 350 (6266) (2015)
1332–1338, publisher: American Association for the Advancement of Sci-
ence Section: Research Article. doi:10.1126/science.aab3050.
URL <https://science.sciencemag.org/content/350/6266/1332>