Abstract

Deep learning has achieved great success in recent years with the aid of advanced neural network structures and large-scale human-annotated datasets. However, it is often costly and difficult to accurately and efficiently annotate large-scale datasets, especially for some specialized domains where fine-grained labels are required. In this setting, coarse labels are much easier to acquire as they do not require expert knowledge. In this work, we propose a contrastive learning method, called masked contrastive learning (MaskCon) to address the under-explored problem setting, where we learn with a coarse-labelled dataset in order to address a finer labelling problem. More specifically, within the contrastive learning framework, for each sample our method generates soft-labels with the aid of coarse labels against other samples and another augmented view of the sample in question. By contrast to self-supervised contrastive learning where only the sample’s augmentations are considered hard positives, and in supervised contrastive learning where only samples with the same coarse labels are considered hard positives, we propose soft labels based on sample distances, that are masked by the coarse labels. This allows us to utilize both inter-sample relations and coarse labels. We demonstrate that our method can obtain as special cases many existing state-of-the-art works and that it provides tighter bounds on the generalization error.

Experimentally, our method achieves significant improvement over the current state-of-the-art in various datasets, including CIFAR10, CIFAR100, ImageNet-1K, Stanford Online Products and Stanford Cars196 datasets. Code and annotations are available at https://github.com/MrChenFeng/MaskCon_CVPR2023.

1. Introduction

Supervised learning with deep neural networks has achieved great success in various computer vision tasks such as image classification, action detection and object localization. However, the success of supervised learning relies on large-scale and high-quality human-annotated datasets, whose annotations are time-consuming and labour-intensive to produce. To avoid such reliance, various learning frameworks have been proposed and investigated: Self-supervised learning aims to learn meaningful representations with heuristic proxy visual tasks, such as rotation prediction [16] and the more prevalent instance discrimination task, the latter, being widely applied in self-supervised contrastive learning framework; semi-supervised learning usually considers a dataset for which only a small part is annotated – within this setting, pseudo labelling methods [24] and consistency regularization techniques [1, 29] are typically used; Moreover, learning using more accessible but noisy data, such as web-crawled data, has also received increasing attention [13, 25].

In this work, we consider an under-explored problem setting aiming at reducing the annotation effort – learning fine-grained representations with a coarsely-labelled dataset. Specifically, we learn with a dataset that is fully labeled, albeit at a coarser granularity than we are interested in (i.e., that of the test set). Compared to fine-grained labels, coarse labels are often significantly easier to obtain, especially in some of the more specialized domains, such as the recognition and classification of medical pathology images. As a simple example, for the task of differentiation between different pets, we need a knowledgeable cat lover to distinguish between ‘British short’ and ‘Siamese’.
but even a child annotator may help to discriminate between ‘cat’ and ‘non-cat’ (Fig. 1). Unfortunately, learning with a coarse labelled dataset has been less investigated compared to other weakly supervised learning paradigms. Recently, Bukchi et al. [2] investigate on learning with coarse labels in the few-shot setting. More closely related to us, Grafit [31] proposes a multi-task framework by a weighted combination of self-supervised contrastive learning and supervised contrastive learning cost; Similarly, CoIns [37] uses both a self-supervised contrastive learning cost and a supervised learning cross-entropy loss. Both works combine a fully supervised learning cost (cross entropy or contrastive) with a self-supervised contrastive loss – these works are the main ones with which we compare.

Differently than them, instead of using self-supervised contrastive learning as an auxiliary task, we propose a novel learning scheme, namely **Masked Contrastive Learning (MaskCon)**. Our method aims to learn by considering inter-sample relations of each sample with other samples in the dataset. Specifically, we always consider the relation to oneself as confidently positive. To estimate the relations to other samples, we derive soft labels by contrasting an augmented view of the sample in question with other samples, and further improve it by utilizing the mask generated based on the coarse labels. Our approach generates soft inter-sample relations that can more accurately estimate fine inter-sample relations compared to the baseline methods (Fig. 1). Efficiently and effectively, our method achieves significant improvements over the state-of-the-art in various datasets, including CIFARtoy, CIFAR100, ImageNet-1K and more challenging fine-grained datasets Stanford Online Products and Stanford Cars196.

## 2. Related works

In this section, we first briefly review the contrastive learning works which are the basis of our method and then, we briefly introduce the state-of-the-art works in related problem settings.

### Self-supervised contrastive learning

Contrastive learning has recently emerged as a powerful method for representation learning without labels. In the context of self-supervision, contrastive learning has been used for instance discrimination, producing models whose performance rivals that of fully supervised training. The objective of contrastive instance discrimination is to minimize the distance between transformed views of the same sample in the feature space, while maximizing their distance from views of other samples. Notable methods in this area are SimCLR [5] and MoCo [18], as well as other variants including SwAV [4], which leverages clustering to identify prototypes for the contrastive learning objective and [33], that reformulates the instance discrimination contrastive objective in the form of a triplet loss. These methods have been found to produce models that, requiring only limited amounts of annotated data for fine-tuning on a particular task or domain, achieve performance that reaches or even surpasses that of supervised models [6, 11]. Furthermore, self-supervised training has been found to compare positively with supervised training in other respects, such as producing models that perform better in the context of continual learning [14].

However, self-supervised contrastive learning based on instance discrimination task usually suffers from the ‘false negative’ problem – that samples from the same class should not be considered as negative. For example, pictures of different cats should not be considered completely negative to each other. To this end, heuristically modifying the inter-sample relations have been widely applied in self-supervised learning [10,12,22,34], intuitively similar to us. Many of the state-of-the-art methods can be considered as special cases of our method by the adjustment of hyperparameters (Sec. 3.2).

### Combining contrastive learning with supervised learning

The success of self-supervised contrastive learning has led to increasing attention on how to integrate contrastive learning into existing supervised learning paradigms. One approach is supervised contrastive learning, which adapts contrastive learning to the fully supervised setting. Supervised contrastive learning was first introduced in [17,28] under the name of neighbourhood components analysis (NCA). Recently, supervised contrastive learning has regained interest due to the significant progress made in [20] by employing more advanced network structures and image augmentations. Compared to standard supervised learning using cross-entropy loss, supervised contrastive learning exhibits superior performance in hard sample mining [20].


### Hierarchical image classification & Unsupervised clustering

Hierarchical image classification methods aim to learn better representations with datasets having hierarchically structured labels. To utilize information from different levels of labels, Zhang et al. [39] propose a multi-losses
contrastive learning framework with each loss considering a specific level of labels. Unsupervised clustering methods often have the same problem setting but different goals than self-supervised learning, that is, to recover and identify the ground-truth semantic labels of each sample [3,32]. DeepCluster [3] simultaneously learns the clustering assignments and a parametric classifier, by feeding the clustering assignments as pseudo labels to the classifier. SCAN [32] inherits the idea of DeepClustering while augmenting extra self-labelling and neighbor mining techniques.

3. Methodology

We consider the problem of learning when the labels of the training data and the labels of the test data are inconsistent in granularity, and in particular when the labels of the training set are coarser. More specifically, let us denote with \( X = \{x_i\}_{i=1}^N \) an i.i.d sampled train dataset with annotated coarse labels \( Y = \{y_i \in \{0, 1\}^M\}_{i=1}^N \) subject to \( \sum_{m=1}^M y_{im} = 1 \). Here, \( N \) denotes the number of samples and \( M \) denotes the number of coarse classes. Let us also define a finely labelled set \( Y' = \{y'_i \in \{0, 1\}^{M'}\}_{i=1}^N \) subject to \( \sum_{k=1}^{M'} y'_{ik} = 1 \), where \( M' \) denotes the number of fine classes. Typically, \( M' > M \).

3.1. Preliminaries

We first quickly review the common learning paradigms utilized in the baseline methods and our method.

3.1.1 Supervised learning

In most supervised learning frameworks, the objective is the minimization of the below empirical risk:

\[
R(f, g) = \sum_{i=1}^N L_{ce}(x_i, y_i; f, g),
\]

where \( f \) denotes the feature encoder, and \( g \) denotes the classifier head (usually a single fully-connected layer).

For brevity, we denote \( f_i \triangleq f(x_i), g_i \triangleq g(f(x_i)) \) and \( p_i \triangleq \text{softmax}(g_i) \) as the feature, logit and prediction of a sample view \( x_i \), respectively. Here,

\[
L_{ce}(x_i, y_i; f, g) = - \sum_{m=1}^M y_{im} \log p_{im}^g
\]

is the cross-entropy loss function – the default loss function for classification problems.

3.1.2 Contrastive learning

Unlike the common supervised learning model above where a parametric classifier \( g \) is learned based on the semantic labels \( Y \) of the samples \( X \), we can also perform contrastive learning based on the inter-sample relations \( Z = \{z_{ij} \in (0, 1)^N\}_{i=1}^N \), with each entry \( z_{ij} \) depicting the inter-sample relation between \( x_i \) and \( x_j \). Intuitively, \( z_{ij} = 1 \) means that sample \( x_i \) and \( x_j \) generate a strong positive pair. Since each sample may form multiple positive sample pairs, for brevity, we abuse the notation here with \( Z \) denoting also the sample-wise normalized inter-sample relations. To learn such inter-sample relations, instead of a parametric classifier \( g \), the \( f \) encoder is usually followed by a projector \( h \), which is often implemented as an MLP and learned by regularizing the inter-sample relations \( Z \) (Eq. (4)).

More specifically, let us denote by \( h_i \triangleq h(f(x_i)) \) the projection. We first calculate the cosine similarity \( d_i \) between a sample \( x_i \) and the dataset \( H = \{h_n\}_{n=1}^N \):

\[
d_i = [\cos(h_i, h_1), \cos(h_i, h_2), \ldots, \cos(h_i, h_N)],
\]

Let us further define \( q_i \triangleq \text{softmax}(d_i/\tau_0) \), where \( \tau_0 \) is the temperature hyperparameter. Then the following empirical risk will be optimized:

\[
R(f, h) = \sum_{i=1}^N L_{con}(x_i, z_i; f, h),
\]

where the contrastive loss \( L_{con} \) is defined as follows:

\[
L_{con}(x_i, z_i; f, h) = - \sum_{n=1}^N z_{in}^0 \log q_{in}^0.
\]

Self-supervised contrastive learning We first introduce the most prevalent form of contrastive learning currently — self-supervised contrastive learning for learning without labels. Since there are no annotated labels, we usually set \( Z_{self} \) by considering each sample having only itself as positive:

\[
z_{ij}^{self} = \begin{cases} 
1, & \text{if } i = j \\
0, & \text{if } i \neq j 
\end{cases}
\]

Typically, we aim at maximizing the relations between different augmented views of the same sample while minimizing the relations between augmented views of different samples. This is also widely known as instance discrimination task. We denote such self-supervised contrastive loss as \( L_{selfcon} \).

Supervised contrastive learning We can easily transition from self-supervised contrastive learning to supervised contrastive learning, by simply changing the inter-sample relations \( Z^{sup} \) for Eq. (4) and Eq. (5). More specifically, as in [17,20,28], samples from the same semantic classes are considered positive pairs,

\[
z_{ij}^{sup} = \begin{cases} 
1, & \text{if } y_i = y_j \\
0, & \text{if } y_i \neq y_j 
\end{cases}
\]

in addition to the sample itself (and its augmented views) as in self-supervised contrastive learning. We denote the supervised contrastive loss as \( L_{supcon} \).
Entropy and CoIns [37] can be expressed as follows:

\[ L \]

Previously mentioned loss functions in Sec. 3.1 (under-clustering). Formally, given our definitions of the pre-

The supervised contrastive loss and CoIns [37] with the the

the supervised cross-entropy loss. In both cases, as a result,

However, due to computation and memory constraints, it is often un-

recent works attempt to

they may

they may

Machine learning (they may

This reduces the noise in \( z_i \). Finally, we re-scale the \( z_i' \) with its maximum


to make the closest neighbour as positive as the sample itself and arrive at:

\[ z_{ij}^\text{mask} = \begin{cases} 1, & \text{if } i = j \\ z_{ij}', & \text{if } i \neq j \end{cases} \]

\[ z_{ij} = z_{ij}' / \max(z_i') \]

Compared to \( Z^{\text{supcon}} \), we thus reweight the samples of the same coarse label according to the similarities in the feature space.

We denote the masked contrastive loss as \( L_{\text{maskcon}} \) and, similarly to Grafit and CoIns, we also consider a weighted loss as the final objective:

\[ L = w L_{\text{maskcon}} + (1 - w) L_{selfcon} \]


to the closest sample with the same fine-grained label). However, considering the samples with the same coarse label as posi-

the samples \( j \) that have a different coarse label with the sam-

While it is risky to consider all samples from the same coarse class as positive, we can confidantly identify those samples that do not have the same coarse class as negative. This reduces the noise in \( z_i \). Finally, we re-scale the \( z_i' \) with its maximum

\[ z_{ij} = z_{ij}' / \max(z_i') \]


to make the closest neighbour as positive as the sample itself and arrive at:

\[ z_{ij}^\text{mask} = \begin{cases} 1, & \text{if } i = j \\ z_{ij}' , & \text{if } i \neq j \end{cases} \]

\[ z_{ij} = z_{ij}' / \max(z_i') \]

Compared to \( Z^{\text{supcon}} \), we thus reweight the samples of the same coarse label according to the similarities in the feature space.

We denote the masked contrastive loss as \( L_{\text{maskcon}} \) and, similarly to Grafit and CoIns, we also consider a weighted loss as the final objective:

\[ L = w L_{\text{maskcon}} + (1 - w) L_{selfcon} \]


to the closest sample with the same fine-grained label). However, considering the samples with the same coarse label as posi-

the samples \( j \) that have a different coarse label with the sam-

While it is risky to consider all samples from the same coarse class as positive, we can confidantly identify those samples that do not have the same coarse class as negative. This reduces the noise in \( z_i \). Finally, we re-scale the \( z_i' \) with its maximum

\[ z_{ij} = z_{ij}' / \max(z_i') \]

Compared to \( Z^{\text{supcon}} \), we thus reweight the samples of the same coarse label according to the similarities in the feature space.

We denote the masked contrastive loss as \( L_{\text{maskcon}} \) and, similarly to Grafit and CoIns, we also consider a weighted loss as the final objective:

\[ L = w L_{\text{maskcon}} + (1 - w) L_{selfcon} \]

Relations to SOTA works By adjusting \( w \) and \( \tau \), our method can obtain various existing SOTA methods as special cases. More specifically, for setting \( \tau \) to \( \infty \), our method degenerates to Grafit [31]. Ignoring the mask, i.e., treating all samples as having the same coarse label, our method can also generalize to existing SOTAs in self-supervised learning. For example, NNCLR [10] instead of considering only each sample itself as positive it considers as positive the nearest neighbors as well. Formally, by setting \( w = 1 \) and \( \tau = 0 \), MaskCon degenerates to NNCLR; ASCL [12] proposes to introduce extra positive samples by calculating soft inter-sample relations, adaptively weighted by its own normalized entropy. By setting \( w = \text{norm} \_\text{entropy}(z_i') \), MaskCon degenerates to ASCL.
3.3. Theoretical justification of MaskCon

The fundamental goal of contrastive learning is to identify \( Z \) that suits the problem granularity and accurately represents the relationship between samples. Assuming that there exists an optimal hidden \( \tilde{Z} \), we can define the expected population risk in terms of \( \tilde{Z} \) as shown in Eq. (16).

\[
R_E(f, h) = E_{a_i, [L_{con}(x_i, \hat{z}_i; f, h)]} \tag{16}
\]

Moreover, let us define the empirical risk with respect to \( \tilde{Z} \) as shown in Eq. (17).

\[
\hat{R}(f, h) = \sum_{i=1}^{N} L_{con}(x_i, \hat{z}_i; f, h) \tag{17}
\]

**Theorem 1.** The generalization error bounds of contrastive learning depends on the difference between the inter-sample relations \( Z \) and the optimal hidden inter-sample relations \( \tilde{Z} \), i.e.,

\[
d_L = \frac{1}{n} \sum_{i=1}^{n} \| z_i - \hat{z}_i \|_2 \tag{18}
\]

**Proof.** Let us assume \( L_{con}(x_i, z_i; f, h) \in [a, b] \) and is \( \lambda \)-Lipschitz continuous w.r.t \( z_i \). Let \( |\mathcal{F}| \) and \( |\mathcal{H}| \) be the covering numbers that correspond to the finite hypothesis space of \( f \) and \( g \). We have:

Hoeffding’s inequality

\[
\begin{align*}
|R_E(f, h) - R(f, h)| & \leq \frac{|R_E(f, h) - \hat{R}(f, h)|}{|\mathcal{F}|} \\
& + \frac{|\hat{R}(f, h) - R(f, h)|}{|\mathcal{H}|}
\end{align*}
\]

\[
\leq |a - b| \sqrt{\frac{\log(2|\mathcal{F}| \cdot |\mathcal{H}| / \delta)}{2n}} + \lambda \frac{1}{n} \sum_{i=1}^{n} \| z_i - \hat{z}_i \|_2
\]

with probability at least \( 1 - \delta \).

Based on Theorem 1, it can be concluded that our method performs better than Grafit with high probability, because an appropriate temperature \( \tau \) allows \( Z_{\text{mask}} \) to more accurately estimate the optimal \( \tilde{Z} \) (Fig. 1) in comparison to assigning equal weights to all coarse neighbours.

4. Experiments

In this section, extensive experiments are conducted to validate the findings and effectiveness of the proposed Masked Contrastive Learning (MaskCon) framework. The impact of hyperparameters is thoroughly investigated in Sec. 4.1. In Sec. 4.2, Sec. 4.3 and Sec. 4.4, results are provided for experiments on CIFAR datasets and ImageNet-1K dataset with coarse labels. In Sec. 4.5, we apply our method on more challenging fine-grained datasets, including Stanford Online Products (SOP) dataset [30] and Stanford Cars 196 dataset [23].

To ensure the robustness of the proposed method and exclude the impact of model capacity, the model settings are kept consistent throughout all the experiments, except for the two hyperparameters, \( w \) and \( \tau \), which are ablated in Sec. 4.1. ResNet18 [19] is employed in all experiments, with the initial convolutional layer modified to have a kernel size of 3×3 and a stride of 1, and the initial max-pooling operator removed for CIFAR experiments, to account for the smaller image size (32×32) as suggested in prior works [18]. More implementation and dataset details can be found in Appendix A.

We compare our method with two competing methods: Grafit and CoIns. For a fair comparison, we exhaust the weight \( w \) choices for both methods and report the best achievable results in all experiments. Note that when \( w = 0 \), Grafit and CoIns degenerate to self-supervised contrastive learning denoted as SelfCon. Conversely, when \( w = 1 \), Grafit degenerates to supervised contrastive learning [20] denoted as SupCon, while CoIns degenerates to conventional supervised cross-entropy learning denoted as SupCE. For reference, we also show the results when training with fine labels – this is denoted as SupFINE.

**Evaluation protocol** To evaluate the different methods on the test set with fine labels, we use the recall@K [27] metric widely used in the image retrieval task. Each test image first retrieves top-K nearest neighbours from the test set and receives 1 if there exists at least one image from the same fine class among the top-K nearest neighbours, otherwise 0. Recall@K averages this score over all the test images.

![Figure 2. Recall@1 w.r.t \( w \) and \( \tau \) on CIFAR100 dataset](image)

**4.1. Effect of \( w \) and \( \tau \)**

In this section, we extensively ablate the effect of hyperparameters of MaskCon – weight \( w \) and temperature \( \tau \).
Table 1. Results on CIFARtoy dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>CIFARtoy-goodsplit</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall@1</td>
<td>Recall@2</td>
<td>Recall@5</td>
<td>Recall@10</td>
<td>Recall@1</td>
<td>Recall@2</td>
<td>Recall@5</td>
<td>Recall@10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SelfCon</td>
<td>84.83</td>
<td>91.55</td>
<td>96.35</td>
<td>98.16</td>
<td>84.83</td>
<td>91.55</td>
<td>96.35</td>
<td>98.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grafit</td>
<td>86.61</td>
<td>92.33</td>
<td>97.01</td>
<td>98.38</td>
<td>89.96</td>
<td>94.36</td>
<td>97.61</td>
<td>98.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SupCon</td>
<td>73.84</td>
<td>84.25</td>
<td>92.14</td>
<td>95.46</td>
<td>84.66</td>
<td>90.55</td>
<td>95.15</td>
<td>96.71</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CoIns</td>
<td>86.15</td>
<td>92.76</td>
<td>97.21</td>
<td>98.46</td>
<td>90.55</td>
<td>94.94</td>
<td>97.73</td>
<td>98.71</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SupCE</td>
<td>76.30</td>
<td>85.26</td>
<td>94.65</td>
<td>97.46</td>
<td>87.15</td>
<td>92.85</td>
<td>96.78</td>
<td>98.34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SupFINE</td>
<td>94.11</td>
<td>96.53</td>
<td>98.25</td>
<td>98.96</td>
<td>94.11</td>
<td>96.53</td>
<td>98.25</td>
<td>98.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MaskCon (Ours)</td>
<td>90.28 (13.98↑)</td>
<td>94.04</td>
<td>97.33</td>
<td>98.53</td>
<td>91.56 (4.41↑)</td>
<td>95.23</td>
<td>97.70</td>
<td>98.70</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Results on CIFARtoy dataset.

Specifically, we show the recall@1 results on CIFAR100 dataset in Fig. 2, with $w = \{0, 0.2, 0.5, 0.8, 1.0\}$ and $\tau = \{0, 0.01, 0.05, 0.1, 0.5, \infty\}$.

The results demonstrate a significant improvement in performance upon the inclusion of additional inter-sample relationships ($w \neq 0$) in comparison to self-supervised learning alone ($w = 0$) on the CIFAR dataset. Furthermore, a suitable temperature $\tau$, such as 0.05 or 0.1, consistently yields superior results compared to the simple weighted combination in Grafit ($\tau = \infty$). In general, we recommend initiating the hyperparameter search with $w = 1$ and $\tau = \tau_0$, where $\tau_0$ denotes the temperature for the projection head. Additional results concerning various hyperparameters on other datasets can be found in Appendix B.

4.2. Experiments on CIFARtoy dataset

To simulate the coarse labelling process, we manually generate toy datasets based on the CIFAR10 dataset. Following the problem motivation, a subset of 8 classes is selected from the original 10 classes. Specifically, the classes 'airplane', 'truck', 'automobile', and 'ship' are designated as Class A, while 'horse', 'dog', 'bird', and 'cat' are assigned to Class B. Note that Class A comprises non-organic objects, while Class B comprises animals. Moreover, a bad split of the aforementioned dataset is defined, wherein 'airplane', 'automobile', 'bird', and 'cat' are Class A, and 'horse', 'dog', 'ship', and 'truck' are Class B.

As shown in Tab. 1, our method achieves the best performance in both splits. Moreover, the improvement over the supervised learning with coarse labels (SupCE) on good split (13.98%) is much higher than the bad split (4.41%), which shows the substantial potential of MaskCon in dealing with more realistic coarse-labelled datasets.

In Fig. 3 we visualize the learned features of all test samples in the good split. Clearly, the learned representations with supervised contrastive learning (SupCon) and self-supervised contrastive learning (SelfCon) tend to under-cluster and over-cluster the samples, respectively. By contrast, our method gets more compact and clear clusters, in line with the results when trained with fine labels (SupFINE).

4.3. Experiments on CIFAR100 dataset

The common CIFAR100 dataset has 20 classes of coarse labels in addition to the 100 classes of fine labels, with each coarse class containing five fine-grained classes (500 samples). The results in Tab. 2 show that our method achieves significant improvements over the SOTAs. In particular, it improves the top-1 retrieval precision from 47.25% to 65.52%, approaching the results by the model learned with fine labels (71.13%).

4.4. Experiments on ImageNet-1K dataset

In this section, we evaluate our method on the large-scale ImageNet-1K dataset. For efficiency, we experiment with the downsampled version of ImageNet-1K dataset [7], where each sample was resized to 32x32. Since no official coarse labels exist for ImageNet, we introduce coarse labels. It is worth noting that the better performance on the bad split may seem counter-intuitive at first glance. However, note that a bad split actually makes the labels more informative, as it explicitly helps to discriminate visually similar classes, e.g., 'cat' and 'dog' in different coarse classes in our setting.

In this section, we conduct experiments in a more challenging scenario – fine-grained datasets with only coarse labels. Please note, that in this work, we do not aim to compare with the state-of-the-art works in fine-grained classification, which usually involve more specialized techniques, such as object localization and local feature extraction.

### 4.5. Experiments on fine-grained datasets

In this section, we conduct experiments in a more challenging scenario – fine-grained datasets with only coarse labels. Please note, that in this work, we do not aim to compare with the state-of-the-art works in fine-grained classification, which usually involve more specialized techniques, such as object localization and local feature extraction.

#### 4.5.1 Stanford Online Products (SOP)

The Stanford Online Products (SOP) dataset [30] consists of 22,634 products, with each product having between two and twelve photos from different perspectives, for a total of 120,053 images. In addition, there are 12 coarse classes based on the semantic categories of the products, such as ’bicycle’ and ’kettle’. For the common image retrieval task, the category of the test set is possibly unknown, so we firstly extract a subset with an unknown test category based on the SOP dataset. We then selected the categories with eight or more images. We then split almost equally this subset to obtain the training set and the test set with 25,368 and 25,278 images, consisting of 2,517 and 2,518 classes respectively, denoted as SOP-split1. Moreover, we select all products with twelve images, and then randomly select ten train images and two test images. Thus, we have 1,498 classes with a total of 17,976 images (2,996 test images and 14,980 train images), denoted as SOP-split2.

In Tab. 4 and Tab. 5 we show results on SOP-split1 and SOP-split2, respectively. Similarly, regardless of whether the test class is known or not, our method still achieves significant improvements over state-of-the-art methods. An interesting phenomenon here is, that on the SOP dataset, SelfCon alone is significantly better than its SupCE/SupCon counterpart. We conjecture that this is due to the fact that each fine class of the SOP dataset in fact consists of different views of the same product, and the number of images per fine class is quite small (2 to 12). For such a dataset, the classification of fine classes is closer to instance discrimination rather than the coarse classes classification.

#### 4.5.2 Stanford Cars196

Stanford Cars196 is another widely used fine-grained image classification benchmark. As there are no official annotated coarse labels, we manually group 196 car models into 8 coarse classes based on the common types of cars: ’0: Cab’, ‘1: Sedan’, ‘2: SUV’, ‘3: Convertible’, ‘4: Coupe’, ‘5: Hatchback’, ‘6: Wagon’ and ‘7: Van’.

In Tab. 6, our method again has the best performance. In addition, unlike in the SOP datasets, SupCon and SupCE have an overwhelming advantage over SelfCon. In this dataset, this is not surprising because for each fine class,
there are more images (ranging from 24 images to 68 images) with different colours/backgrounds in the cars dataset, leading to a much higher intra-class variance. Some image retrieval examples can be found in Fig. 4.

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall@1</th>
<th>Recall@2</th>
<th>Recall@5</th>
<th>Recall@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>SelfCon</td>
<td>20.97</td>
<td>28.75</td>
<td>41.44</td>
<td>52.61</td>
</tr>
<tr>
<td>Grafit</td>
<td>42.30</td>
<td>54.79</td>
<td>71.1</td>
<td>81.74</td>
</tr>
<tr>
<td>SupCon</td>
<td>42.30</td>
<td>54.79</td>
<td>71.1</td>
<td>81.74</td>
</tr>
<tr>
<td>Colns</td>
<td>42.77</td>
<td>55.60</td>
<td>72.29</td>
<td>82.53</td>
</tr>
<tr>
<td>SupCE</td>
<td>42.77</td>
<td>55.60</td>
<td>72.29</td>
<td>82.53</td>
</tr>
<tr>
<td>SupFINE</td>
<td>78.09</td>
<td>82.97</td>
<td>86.51</td>
<td>88.04</td>
</tr>
<tr>
<td>MaskCon (Ours)</td>
<td><strong>45.53 (7.16)</strong></td>
<td><strong>58.56</strong></td>
<td><strong>74.36</strong></td>
<td><strong>84.36</strong></td>
</tr>
</tbody>
</table>

Table 6. Results on Stanford Cars196 dataset.

4.5.3 Additional discussion

Compared to the state-of-the-art, our method achieves the best results in all experiments. We would like to note here some of our insights from the results on the CIFAR and the fine-grained ones (including ImageNet-1K). More specifically, we note that on the CIFAR datasets the performance of our method approaches that of the supervised one with fine labels (SupFINE), while on the fine-grained datasets there is a larger gap. We note that the critical question is to consider whether the instance discrimination task is useful for fine-grained classification in different scenarios, since coarse labelling usually does not lead to worse results. Towards that, recent works have shown that the core idea of self-supervised contrastive learning — augmentation invariance may be destructive for fine-grained tasks [8, 36]. For example, the commonly applied colour distortion augmentation will promote the model to be non-sensitive to colour information. However, colour may be the key to discriminate between different breeds of birds. How to adapt the instance discrimination task to the fine-grained coarse-labelled dataset is a future direction we wish to pursue.

5. Conclusion

In this work, we propose a Masked Contrastive learning framework (MaskCon) for learning fine-grained information with coarse-labelled datasets. On the basis of two baseline methods, we utilize coarse labels and the instance discrimination task to better estimate inter-sample relations. We show theoretically that our method can reduce the optimization error bound. Extensive experiments with various hyperparameter settings on multiple benchmarks, including the CIFAR datasets and the more challenging fine-grained classification datasets show that our method achieves consistent and large improvement over the baselines.

Acknowledgments: This work was supported by the EU H2020 AI4Media No. 951911 project.
References


[23] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13), Sydney, Australia, 2013. 5


[28] Ruslan Salakhutdinov and Geoff Hinton. Learning a nonlinear embedding by preserving class neighbourhood struc-


