Adaptive Class Suppression Loss for Long-Tail Object Detection

Tong Wang¹, Yousong Zhu¹, Chaoyang Zhao¹, Wei Zeng⁴,⁵, Jinqiao Wang¹,²,⁶, Ming Tang¹

¹ National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
² School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
³ ObjectEye Inc., Beijing, China
⁴ Peking University, Beijing, China
⁵ Peng Cheng Laboratory, Shenzhen, China
⁶ NEXWISE Co., Ltd., Guangzhou, China

{tong.wang, yousong.zhu, chaoyang.zhao, jqwang, tangm}@nlpr.ia.ac.cn
weizeng@pku.edu.cn

Abstract

To address the problem of long-tail distribution for the large vocabulary object detection task, existing methods usually divide the whole categories into several groups and treat each group with different strategies. These methods bring the following two problems. One is the training inconsistency between adjacent categories of similar sizes, and the other is that the learned model is lack of discrimination for tail categories which are semantically similar to some of the head categories. In this paper, we devise a novel Adaptive Class Suppression Loss (ACSL) to effectively tackle the above problems and improve the detection performance of tail categories. Specifically, we introduce a statistic-free perspective to analyze the long-tail distribution, breaking the limitation of manual grouping. According to this perspective, our ACSL adjusts the suppression gradients for each sample of each class adaptively, ensuring the training consistency and boosting the discrimination for rare categories. Extensive experiments on long-tail datasets LVIS and Open Images show that our ACSL achieves 5.18% and 5.2% improvements with ResNet50-FPN, and sets a new state of the art. Code and models are available at https://github.com/CASIA-IVA-Lab/ACSL.

1. Introduction

With the advent of deep Convolutional Neural Network, researchers have achieved significant progress on object detection task. Many efforts have been paid to refresh the record on classic benchmarks like PASCAL VOC [9] and MS COCO [21]. However, these benchmarks usually have limited quantities of classes and exhibit relatively balanced category distribution. Whereas, in real-world scenarios, data usually comes with a long-tail distribution. A few head classes (frequent classes) contribute most of the training samples, while the huge number of tail classes (rare classes) are under-represented in data. Such extremely imbalanced class distribution proposes new challenges for researchers.

An intuitive solution is to re-balance the data distribution by re-sampling technique [12, 7, 28, 23]. By over-sampling
the tail or under-sampling the head classes, a less imbalanced distribution could be artificially generated. Nevertheless, over-sampling usually brings undesirable over-fitting issues on tail classes. And under-sampling may miss valuable information of head classes. The effectiveness of such techniques is limited when the dataset is extremely imbalanced. To improve the performance of tail classes while avoiding the over-fitting issue, Tan et al. [29] devises Equalization Loss, which argues that the poor performance of tail classes originates from the over-suppression of samples from head classes. Since tail classes only contain few samples, they receive much more negative gradients than positive ones during training, thus they are consistently in a state of being suppressed in most of the training time. In order to prevent tail classifiers from being over-suppressed, Equalization Loss proposes to ignore all negative gradients from head classes. Balanced Group Softmax (BAGS) [18] puts categories with similar numbers of training instances into the same group and computes group-wise softmax cross-entropy loss respectively. BAGS achieves relative balance within each group, thus can effectively ameliorate the domination of the head classes over tail classes.

The above methods can efficaciously reduce the suppression on tail classifiers. However, they need to partition the categories into several groups based on their category frequency prior. Such hard division between head and tail classes brings two problems, namely training inconsistency between adjacent categories and lack of discriminative power for rare categories. As shown in Figure 1, when two categories with similar instance statistics are divided into two different groups, there exists a huge gap between their training strategies. Such training inconsistency may deteriorate the network’s performance. In general, it is sub-optimal to distinguish the head and tail classes by the absolute number of samples. In addition, it frequently happens that two categories with high appearance similarity hold totally different sample frequency for datasets with large category vocabulary. For instance, category “sunglasses” and “eye mask” belong to the head and tail classes, respectively. To prevent the tail classifier “eye mask” from being over-suppressed, the negative gradients from category “sunglasses” is ignored by the classifier of “eye mask”. Nevertheless, this can also reduce the differentiability between these semantically similar cases in feature space, the classifier of “eye mask” has a high chance to misclassify sample “sunglasses” as “eye mask”. For those categories which belong to different groups but have high appearance similarity, it is hard for the network to learn a discriminative feature representation.

Therefore, in this paper, we present a novel Adaptive Class Suppression loss (ACSL) to address the above two problems. The design philosophy is simple and straightforward: We assume all categories are from “tail” group, and adaptively generate suppression gradients for each category according to their current learning status. To be specific, we propose to treat all object categories as scarce categories, regardless of the statistics of per-category instances, thus eliminating the dilemma of manually defining the head and tail. Furthermore, in order to alleviate the insufficient learning and representation, we introduce an Adaptive Class Suppression Loss to adaptively balance the negative gradients between different categories, which effectively boosts the discriminative ability for tail classifiers. On the one hand, the proposed method can exterminate several heuristics and hyper-parameters of data distribution. On the other hand, it can also avoid the problem caused by over-sampling and under-sampling, and ensure the training consistency for all classes and sufficient learning of rare or similar categories. Finally, it yields reliable and significant improvements in detection performance on large-scale benchmarks like LVIS and Open Images.

To sum up, this work makes the following three contributions:

1. We propose a new statistic-free perspective to understand the long-tail distribution, thus significantly avoiding the dilemma of manual hard division.
2. We present a novel adaptive class suppression loss (ACSL) that can effectively prevent the training inconsistency of adjacent categories and improve the discriminative power of rare categories.
3. We conduct comprehensive experiments on long-tail object detection datasets LVIS and Open Images. ACSL achieves 5.18% and 5.2% improvements with ResNet50-FPN on LVIS and OpenImages respectively, which validates its effectiveness.

2. Related Works

General Object Detection Current CNN-based object detectors can be divided into anchor-based and anchor-free detectors based on whether they depend on the anchor heuristic. Classic anchor-based detectors consist of one-stage and two-stage approaches. Two-stage detectors [27, 10, 6, 1, 25, 13, 19, 24] first generate coarse object candidates by a Region Proposal Network (RPN). And then the region features of these proposals will be extracted for accurate classification and bounding box regression. One-stage detectors [22, 37, 20, 3, 36, 35] have a more concise structure. They make predictions on multiple feature maps directly without proposal generation process, thus enjoying higher computational efficiency.

Anchor-free pipelines abandon the anchor mechanism. They first locate several predefined or self-learned keypoints and then group them to final detections. CornerNet [16, 17], ExtremeNet [38] and CenterNet [8] etc. represent one object as several predefined keypoints and detect
by predicting these keypoints. RepPoints [34] proposes a finer representation of objects as a set of sample points and learns to arrange them in a manner that bounds the spatial extent of an object. FCOS [30] solves object detection in a per-pixel prediction fashion. It predicts four distances for each location and filters those unreliable detections with novel centerness scores. They all perform well on balanced datasets. However, directly applying them to long-tail datasets achieves inferior performance due to the imbalanced data distribution. Thus, our intention is to improve the detectors’ performance on long-tail datasets.

Long-tail Recognition Re-sampling strategy is a typical technique for imbalanced datasets. Class-aware sampling [28] and Repeat factor sampling [11] share the same design philosophy that they aim to balance the data distribution by adopting different sampling frequencies for different categories. Re-weighting [33, 14] is another widely used method which works by assigning weights for different training samples to emphasize the training of tail samples. Cui et al. [5] further proposes Class balanced loss which first calculates the effective number of samples for each category and uses the effective numbers to re-balance the loss.

In addition, some works design specialized loss functions or training strategies to tackle the imbalanced issue. Equalization Loss [29] ignores the negative gradients from head samples to prevent the tail classifiers from being over-suppressed. Cao et al. [2] proposes label-distribution-aware margin loss motivated by minimizing a margin-based generalization bound. Balanced Group Softmax [18] puts categories with similar sample numbers to one group and applies softmax within each group. Balanced Meta-Softmax [26] devises an unbiased extension of softmax to accommodate the label distribution shift between training and testing. SimCal [31] proposes a simple calibration framework to more effectively alleviate classification head bias with a bi-level class balanced sampling approach. Kang et al. [15] design a decoupling training schema, which first learns the representations and classifier jointly, then obtains a balanced classifier by re-training the classifier with class-balanced sampling.

The above methods can relieve the imbalanced training to a great extent. However, they all depend on the class distribution which brings inconvenience when we apply them to new long-tail datasets. In this paper, we devise a more general loss function which does not rely on the category frequency prior and can also handle the long-tail datasets well.

### 3. Adaptive Class Suppression Loss

In this section, we start by analyzing the limitations of group-based methods. Then, we introduce our proposed method.

<table>
<thead>
<tr>
<th>Groups</th>
<th>$\text{mAP}$</th>
<th>$\text{AP}_r$</th>
<th>$\text{AP}_c$</th>
<th>$\text{AP}_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.5)(5,∞)</td>
<td>22.74</td>
<td>6.83</td>
<td>22.14</td>
<td>29.83</td>
</tr>
<tr>
<td>(0.50)(50,∞)</td>
<td>25.30</td>
<td>15.11</td>
<td>24.99</td>
<td>29.77</td>
</tr>
<tr>
<td>(0.500)(500,∞)</td>
<td>25.66</td>
<td>13.19</td>
<td>25.98</td>
<td>30.25</td>
</tr>
<tr>
<td>(0.5000)(5000,∞)</td>
<td>23.89</td>
<td>8.27</td>
<td>23.87</td>
<td>30.16</td>
</tr>
</tbody>
</table>

Adaptive Class Suppression Loss in detail and summary its advantages compared with previous methods.

### 3.1. Limitations of Group-based Methods

To pursue a balanced training for tail classifiers, several works propose to divide the whole categories into several groups based on the category frequency prior and adopt different training strategies for different groups. For instance, Equalization Loss [29] ignores the negative gradients from samples of head categories to prevent the tail classifiers from being dominated by negative signals. Balanced Group Softmax [18] first puts categories with similar sample numbers into the same group and applies softmax cross-entropy loss within each group.

The group-based methods show their ability to improve the performance of tail classifiers. For these methods, how to properly divide all categories into different groups is of vital importance. We empirically find that the quality of grouping directly impacts the performance. We choose BAGS (one of the state-of-the-art group-based methods for long-tail object detection) with ResNet50-FPN as an example to illustrate the importance of proper grouping in LVIS dataset. For clarity, we divide all categories into 2 groups rather than the default 4 groups. We utilize different group partition strategies to train the detectors and report the $m\text{AP}$, $\text{AP}_r$ ($r$ for rare categories), $\text{AP}_c$ ($c$ for common) and $\text{AP}_f$ ($f$ for frequent) in Table 1. As we can see from this table, the dividing line need to be settled in a proper range (50’500) to achieve satisfactory performance. Setting it to a too large or small number all deteriorate the performance.

Since the proper grouping is the precondition of good performance, it is necessary to divide groups based on the data distribution when applying group-based methods to other long-tail datasets. Nevertheless, different datasets usually have various data distributions. A good group partition for one dataset may be suboptimal for other datasets. As shown in Figure 2, models can obtain the best performance when the dividing line is set to 500 in LVIS. However, it is not a good partition for Open Images dataset since it has fewer categories and each category has more samples. Finding a proper group partition strategy for a new dataset could be laborious and time-consuming, which will limit the application of group-based methods in real-world scenarios. We naturally ask a question: can we devise a more general method that can be directly applied to other long-tail datasets?
3.2. Formulation of ACSL

Based on the above analysis, we believe a satisfactory loss function for long-tail datasets should have the following two properties:

(i) **The tail classifiers should be protected from being over-suppressed by the overwhelming head samples. In the meantime, negative gradients for easily confused categories should be preserved for discriminative learning.**

(ii) **The loss function should better not depend on the label distribution prior so that it can be seamlessly applied to diverse datasets without re-calculating the new category distribution statistics.**

Giving a sample $x_s$ belonging to the head category $k$, its label is a one-hot vector $Y$, in which $y_k$ equals to 1 and $y_i (i \neq k)$ is set to 0. $z_i$ represents the network output logit of category $i$. By applying the sigmoid function to the logit $z_i$ (Equation (1)), we can obtain $p_i$, the probability of current sample belonging to category $i$. The binary cross-entropy loss is formulated as Equation (2). Then we can derive the gradient of the loss function with respect to logit $z_i$ as in Equation (4). For tail category $i \neq k$, BCE generates negative suppression gradients to force the classifier $i$ to output low confidence. Such suppression gradients is beneficial to some extent. However, the excessive suppression gradients derived from the head categories will seriously hinder the positive activation of tail categories.

\[
p_i = \frac{1}{1 + e^{-z_i}}, \quad (1)
\]

\[
L_{BCE}(x_s) = -\sum_{i=1}^{C} \log(\hat{p}_i) \quad (2)
\]

where,

\[
\hat{p}_i = \begin{cases} 
  p_i, & \text{if } i = k \\
  1 - p_i, & \text{if } i \neq k
\end{cases} \quad (3)
\]

\[
\frac{\partial L_{BCE}}{\partial z_i} = \begin{cases} 
  p_i - 1, & \text{if } i = k \\
  p_i, & \text{if } i \neq k
\end{cases} \quad (4)
\]

Therefore, we devise a novel Adaptive Class Suppression Loss to protect the training of tail classifiers and adaptively choose which categories should be suppressed based on the learning status. As shown in Equation (5), we multiply a binary weight term $w_i$ to the loss term $-\log(\hat{p}_i)$ for category $i$. For category $k$, $w_i$ is set to 1 since the current sample belongs to category $k$. For other categories $i (i \neq k)$, $w_i$ controls whether the network applies suppression on category $i$. Here, we utilize the output confidence $p_i$ as a signal to determine whether to suppress category $i$. If $p_i$ is larger than the predefined threshold $\xi$, that means the network is confused between category $i$ and $k$. Hence, we set $w_i$ to 1 to perform discriminative learning. Otherwise, $w_i$ will be set to 0 to avoid numerous unnecessary negative suppression. Rather than depending on the category distribution statistics, our proposed ACSL only relies on the network output confidences which saves the efforts of finding optimal category statistics related hyper-parameters when switching to a new dataset. The formulation is defined as follows:

\[
L_{ACSL}(x_s) = -\sum_{i=1}^{C} w_i \log(\hat{p}_i) \quad (5)
\]
where,

\[ w_i = \begin{cases} 
1, & \text{if } i = k \\
1, & \text{if } i \neq k \text{ and } p_i \geq \xi \\
0, & \text{if } i \neq k \text{ and } p_i < \xi 
\end{cases} \quad (6) \]

The gradient of the loss function with respect to \( z_i \) can be derived as Equation (7).

\[ \frac{\partial L_{ACSL}}{\partial z_i} = \begin{cases} 
-p_i - 1, & \text{if } i = k \\
-w_ip_i, & \text{if } i \neq k 
\end{cases} \quad (7) \]

As shown in Figure 3, we give a simple illustration to better understand how ACSL works. For sigmoid cross-entropy loss (Figure 3(a)), it does not consider the imbalanced class distribution, and directly generate negative gradients for all other categories (the weight vector is filled with 1), thus leading to the tail categories receiving a large number of suppression information continuously, which severely reduces the discrimination ability of tail classifiers. For the proposed ACSL, the network can adaptively generate suppression gradients for tail categories. As shown in Figure 3(b), the network simultaneously yields high confidence 0.9 for category “D”, which means the category “A” is semantically similar to the category “D”. Therefore, it is necessary to generate negative suppression for category “D”, but not for the remaining tail categories with low confidences.

### 3.3. Advantages over Previous Methods

Compared with previous methods, the proposed ACSL has the following three characters that makes it more suitable for the long-tail datasets:

* **ACSL takes the network learning status into consideration.** Previous methods [7, 28, 29, 18] only consider the label distribution, while ignore the network learning status. For instance, state-of-the-art Equalization Loss [29] calculates a binary weight for each category based on the sample numbers. Once the dataset is given, the weight will be determined and will not be changed during the whole training process. ACSL adaptively choose which categories to suppress based on the output confidences, which takes the network learning status into consideration, resulting in a more efficient learning process.

* **ACSL works in a more fine-grained sample level.** Previous methods perform an identical operation on samples from the same category. For example, Class balanced loss [7] assigns the same weights for samples from one category. Equalization loss [29] generates the same weight masks for two samples if they all belong to the same category. These methods all ignore the diversity of variant samples. In contrast, ACSL calculates the category weights for a sample based on its output confidences. The generated category weights for samples from the same category might be different. Thus, ACSL can control the training of each classifier in a more accurate way.

* **ACSL does not depend on the class distribution.** The label distribution is indispensable for previous methods. They need to know the class distribution in advance to design sampling strategy [28, 11], determine the weights for samples [7] or divide categories into groups [18, 29]. This is inefficient since the strategies must be re-designed based on the new data distribution when we want to apply these methods to new long-tail datasets. In comparison, ACSL does not require the category frequency prior, which means it can be applied to new long-tail datasets seamlessly. We empirically find that ACSL works well on LVIS and Open Images under the same hyper-parameter setting.

### 4. Experiments on LVIS

#### 4.1. Dataset and Setup

To validate the effectiveness of the proposed ACSL, we conduct comprehensive experiments on the long-tail Large Vocabulary Instance Segmentation (LVIS) dataset [11]. In this work, we use the LVIS-v0.5, which contains 1230 categories with both bounding box and instance mask annotations. We train the models with 57k \textit{train} images and report the accuracy on 5k \textit{val} images. LVIS divides all categories into 3 groups based on the number of images that
contain those categories: frequent (>100 images), common (11-100 images) and rare (<10 images). We use the official evaluation API\(^1\) to evaluate all the models. Besides the official metrics \(mAP\), the \(AP_c\) (AP for rare classes), \(AP_r\) (AP for common classes) and \(AP_f\) (AP for frequent classes) are also reported.

### 4.2. Implementation Details

We choose the classic two-stage detector Faster R-CNN with FPN structure as the baseline. The training image is resized such that the shorter edge is 800 pixels and the longer edge is no more than 1333. When training, we use 8 GPUs with a total batch size 16. The optimizer is stochastic gradient descent (SGD) with momentum 0.9 and weight decay 0.0001. The initial learning rate is 0.02 with 500 iterations’ warm up. For 1x training schedule, learning rate decays to 0.002 and 0.0002 at epoch 8 and 11, respectively. The training stops at the end of epoch 12. When testing, we first apply Non-Maximum Suppression with IoU threshold 0.5 to remove duplicates. Then, the top 300 detections will be selected as the final results. Other hyper-parameter settings like anchor scales and anchor ratios follow the default settings in MMDetection [4].

When we conduct experiments with ACSL or other specialized loss functions designed for long-tail datasets, we focus on the classification subnet of Faster R-CNN and replace the softmax cross-entropy loss with these specialized loss functions. Inspired by [15], we decouple the training of feature representation and the classifier. We train a naive Faster R-CNN detector in the first stage and fine-tune the classifier in the second stage. In object detection, the background samples are negative samples for all categories. We adopt a simple strategy to chase a relative balance for each classifier. Since the ratio between the image numbers of rare, common and frequent categories is approximately 1:10:100, we randomly choose 1%, 10% background samples for rare and common categories, respectively.

#### 4.3. Ablation Study

We take the Faster R-CNN with ResNet50-FPN backbone as the baseline model. As we first train a naive Faster R-CNN detector for 12 epochs in the first stage and then fine-tune the classifier with ACSL for another 12 epochs, the total training epoch is 24. We also report the results of the baseline model trained with 24 epochs for a fair comparison. As the results in Table 2, the baseline (1x) model has a relatively high precision on frequent categories but an extremely low accuracy on rare categories, merely 4.3%. More training iterations bring benefits to rare and common categories and lift the overall \(mAP\) by 1.1%. Even so, the performance of rare categories is still unsatisfactory, which shows the bottleneck of the baseline model lies in the tail classes.

ACSL introduces a hyper-parameter \(\xi\) to define the easily confused categories. It is a trade-off between relieving over-suppression on tail classes and chasing discriminative learning. A small \(\xi\) means that most of the categories will be suppressed, which will suppress too much on tail categories. However, for an extremely large \(\xi\), the network will only suppress categories with extremely high confidences while ignore most of the other categories, thus will weaken the classifier’s discriminative power. To explore how \(\xi\) influences the performance, we conduct experiments with several different values and report the results in Table 2. As shown in this table, setting \(\xi\) to 0.01 improves the precision but the improvements are limited. As \(\xi\) grows larger, the \(mAP\) and \(AP_c\) also increase which proves that properly increasing \(\xi\) relieves the suppression on tail classifiers. When \(\xi\) increases to 0.7, the model achieves the best performance. The overall \(mAP\) is 26.36%, surpassing the baseline model by considerable 5.18%. Nevertheless, continuing to increase its value deteriorates the performance since it will weaken the discriminative power of the classifier. We empirically find \(\xi = 0.7\) works best under current setting.

#### 4.4. Generalization on Stronger Models

To verify the generalization of our method, we further conduct experiments on stronger backbones and de-

---

\(^1\) https://github.com/lvis-dataset/lvis-api
Table 4: Comparison with state-of-the-art methods on LVIS-v0.5 val dataset. **Bold** numbers denote the best results among all models. “ms” means multi-scale testing.

<table>
<thead>
<tr>
<th>Methods</th>
<th>backbone</th>
<th>mAP</th>
<th>AP$_r$</th>
<th>AP$_c$</th>
<th>AP$_f$</th>
<th>AP@0.5</th>
<th>AP@0.75</th>
<th>AP$_r$</th>
<th>AP$_c$</th>
<th>AP$_f$</th>
<th>AP$_r$</th>
<th>AP$_c$</th>
<th>AP$_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBL [5]</td>
<td></td>
<td>23.9</td>
<td>11.4</td>
<td>23.8</td>
<td>27.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDAM [2]</td>
<td></td>
<td>23.9</td>
<td>11.4</td>
<td>23.8</td>
<td>27.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RFS [11]</td>
<td></td>
<td>24.9</td>
<td>15.4</td>
<td>25.3</td>
<td>26.3</td>
<td>41.6</td>
<td>25.8</td>
<td>19.8</td>
<td>30.6</td>
<td>37.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SimCal [31]</td>
<td></td>
<td>23.4</td>
<td>16.4</td>
<td>22.5</td>
<td>27.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EQL [29]</td>
<td>ResNet-50</td>
<td>25.06</td>
<td>11.92</td>
<td>25.98</td>
<td>29.14</td>
<td>40.14</td>
<td>27.30</td>
<td>20.08</td>
<td>31.50</td>
<td>38.67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet-101</td>
<td>ResNet-101</td>
<td>26.05</td>
<td>11.45</td>
<td>27.14</td>
<td>30.51</td>
<td>41.30</td>
<td>27.83</td>
<td>20.35</td>
<td>33.73</td>
<td>40.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNeXt-101-64x4d</td>
<td>ResNeXt-101-64x4d</td>
<td>28.04</td>
<td>15.03</td>
<td>29.14</td>
<td>31.87</td>
<td>44.06</td>
<td>30.07</td>
<td>22.19</td>
<td>34.52</td>
<td>42.97</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BAGS [18]</td>
<td>ResNet-50</td>
<td>25.96</td>
<td>17.65</td>
<td>25.75</td>
<td>29.54</td>
<td>43.58</td>
<td>27.15</td>
<td>20.26</td>
<td>32.81</td>
<td>40.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet-101</td>
<td>ResNet-101</td>
<td>26.39</td>
<td>16.80</td>
<td>25.82</td>
<td>30.93</td>
<td>43.44</td>
<td>27.63</td>
<td>20.29</td>
<td>34.39</td>
<td>41.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNeXt-101-64x4d</td>
<td>ResNeXt-101-64x4d</td>
<td>27.83</td>
<td>18.78</td>
<td>27.32</td>
<td>32.07</td>
<td>45.83</td>
<td>28.99</td>
<td>21.92</td>
<td>35.65</td>
<td>43.11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACSL (Ours)</td>
<td>ResNet-50</td>
<td>26.36</td>
<td>18.64</td>
<td>26.41</td>
<td>29.37</td>
<td>42.38</td>
<td>28.63</td>
<td>20.43</td>
<td>33.11</td>
<td>40.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet-101</td>
<td>ResNet-101</td>
<td>27.49</td>
<td>19.25</td>
<td>27.60</td>
<td>30.65</td>
<td>43.45</td>
<td>29.69</td>
<td>21.11</td>
<td>34.96</td>
<td>42.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNeXt-101-64x4d</td>
<td>ResNeXt-101-64x4d</td>
<td>28.93</td>
<td><strong>21.78</strong></td>
<td>28.98</td>
<td>31.72</td>
<td>45.54</td>
<td>31.19</td>
<td>22.16</td>
<td>35.81</td>
<td>43.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet-50 (ms)</td>
<td>ResNet-50</td>
<td>27.24</td>
<td>17.86</td>
<td>27.42</td>
<td>30.76</td>
<td>44.46</td>
<td>28.54</td>
<td>20.96</td>
<td>34.40</td>
<td>41.68</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet-101 (ms)</td>
<td>ResNet-101</td>
<td>28.23</td>
<td>17.42</td>
<td>28.40</td>
<td>32.32</td>
<td>44.73</td>
<td>30.13</td>
<td>21.86</td>
<td>35.43</td>
<td>44.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNeXt-101-64x4d (ms)</td>
<td>ResNeXt-101-64x4d (ms)</td>
<td><strong>29.47</strong></td>
<td>20.30</td>
<td><strong>29.45</strong></td>
<td>33.15</td>
<td><strong>46.82</strong></td>
<td><strong>31.55</strong></td>
<td><strong>22.52</strong></td>
<td><strong>37.32</strong></td>
<td><strong>45.51</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ACSL still exists when we use a larger backbone ResNeXt-101. The mAP of ACSL is 4.23% higher than the baseline model for ResNeXt-101. Moreover, the utilization of ACSL is not limited to a certain type of detector. It can be easily combined with other detectors such as Cascade R-CNN. For Cascade R-CNN, we replace the softmax cross-entropy loss with ACSL on all 3 heads and fine-tune the classifiers’ weights in the second training stage. With ResNet-101 Cascade R-CNN detector, ACSL achieves 29.71% mAP, surpassing baseline model by 4.57% mAP. When using larger backbone ResNeXt101, the overall mAP can be further pushed to 31.47%, outperforming the baseline by a significant 4.33%. It is worth noticing that ACSL obtains the best performance on the overall mAP, AP$_r$ and AP$_c$, which further proves ACSL’s ability to tackle long-tail datasets.

4.5. Performance Analysis

In order to have a more intuitive sense of how ACSL influences the network’s performance, we visualize the performance of baseline and ACSL on different categories in Figure 4. Figure 4(a) shows the AP on frequent categories. The two curves almost overlap with each other which demonstrates that ACSL does not harm the training of head classifiers. For common categories (Figure 4(b)), ACSL begins to show its advantages over baseline model. The precision curve of ACSL almost covers that of the baseline model, which shows that ACSL brings moderate improvements for common categories. As the decreasing of categories’ sample numbers, the advantages of ACSL becomes more significant. As showed in Figure 4(c), ACSL outperforms baseline by a landslide. The integral area of the orange curve (ACSL) is much larger than that of the blue one (baseline). These three figures indicate that ACSL is able to improve the performance of tail categories without sacrificing the precision of head classes.

4.6. Comparison with State of the Arts

In this section, we compare the performance of the proposed ACSL with other state-of-the-art methods and report the results in Table 4. All the models adopt Faster R-CNN with FPN structure. As shown in the table, without any bells and whistles, our single ResNet-50 model achieves 26.36% mAP. It surpasses other competitive methods, including Equalization Loss (25.06% mAP) and BAGS (25.96% mAP). And we do observe that the performance gain comes from the tail classes (rare and common categories). Among all the single ResNet-50 models, ACSL obtains the best performance on rare (18.64%) and common categories (26.41%). In the meantime, it does not sacrifice the precision of head classes in performance on the overall categories R-CNN detector, ACSL achieves 29.71% mAP, surpassing baseline model by 4.57% mAP. When using larger backbone ResNeXt-101, the overall mAP can be further pushed to 31.47%, outperforming the baseline by a significant 4.33%. It is worth noticing that ACSL obtains the best performance on the overall mAP, AP$_r$ and AP$_c$, which further proves ACSL’s ability to tackle long-tail datasets.

4.5. Performance Analysis

In order to have a more intuitive sense of how ACSL influences the network’s performance, we visualize the performance of baseline and ACSL on different categories in Figure 4. Figure 4(a) shows the AP on frequent categories. The two curves almost overlap with each other which demonstrates that ACSL does not harm the training of head classifiers. For common categories (Figure 4(b)), ACSL begins to show its advantages over baseline model. The precision curve of ACSL almost covers that of the baseline model, which shows that ACSL brings moderate improvements for common categories. As the decreasing of categories’ sample numbers, the advantages of ACSL becomes more significant. As showed in Figure 4(c), ACSL outperforms baseline by a landslide. The integral area of the orange curve (ACSL) is much larger than that of the blue one (baseline). These three figures indicate that ACSL is able to improve the performance of tail categories without sacrificing the precision of head classes.

4.6. Comparison with State of the Arts

In this section, we compare the performance of the proposed ACSL with other state-of-the-art methods and report the results in Table 4. All the models adopt Faster R-CNN with FPN structure. As shown in the table, without any bells and whistles, our single ResNet-50 model achieves 26.36% mAP. It surpasses other competitive methods, including Equalization Loss (25.06% mAP) and BAGS (25.96% mAP). And we do observe that the performance gain comes from the tail classes (rare and common categories). Among all the single ResNet-50 models, ACSL obtains the best performance on rare (18.64%) and common categories (26.41%). In the meantime, it does not sacrifice the precision of head classes in performance on the overall categories.
provements on tail classes. The observations above exhibit the ability of ACSL to tackle the class imbalance problem in long-tail datasets.

The advantages of ACSL becomes more significant when it is combined with larger backbones. With ResNet-101 backbone, ACSL achieves 27.49% mAP, outperforming Equalization Loss and BAGS by 1.44% and 1.1% mAP, respectively. For stronger backbone ResNeXt-101-64x4d, ACSL obtains a 28.93% mAP, surpassing the counterparts of Equalization Loss (28.04% mAP) and BAGS (27.83% mAP). When applying multi-scale testing strategy, the performance can be further improved. Finally, our best model achieves 29.47% mAP.

5. Experiments on Open Images

Open Images V5 is the largest existing object detection dataset which contains a total of 16M bounding boxes for 600 object classes on 1.9M images. For such a large scale dataset, it is hard to maintain a relatively balanced class distribution. In fact, it suffers from severe class imbalance. For instance, the most frequent category “Man” contains 378077 images, which is 126 thousand times of the rarest category “Paper cutter” (3 images). The whole dataset consists of three parts: train, validation and test. All models are trained with 1.7M train images and evaluated on validation dataset. When performing evaluation, we use the official evaluation code for Open Images which evaluates only at IoU threshold 0.5. To reduce the training time on such a large dataset, we use the large batch training framework LargeDet [32] to train our models.

To testify the effectiveness of ACSL, we conduct experiments with various backbones including ResNet50, ResNet101 and ResNet152. The hyper-parameter $\xi$ of ACSL is set to 0.7, the same value as in LVIS. Since objects in Open Images have multiple labels, we train the models under multiple label setting. The experimental results are summarized in Table 5. As shown in this table, the models trained with ACSL outperform the baseline models by a large margin. For ResNet50-FPN detector, ACSL achieves 60.3% mAP, 5.2% higher than the baseline. For larger backbone ResNet152, our model obtains 62.8% mAP, surpassing the baseline by 5.4% mAP. To validate that the performance gains mainly come from the rare categories, we extract the precisions of some rare categories and list them in Table 6. We can observe that ACSL achieves remarkable performance improvements for rare categories. For rare category “Face powder” (“Fac” for short), the performance gain is 63.1% mAP. As shown in this table, ACSL brings considerable improvements to other rare categories, leading to higher overall precision.

We also compare our method with other methods on Open Images dataset. The results are summarized in Table 7. From Table 7, we can observe the performance advantage of our method with the same backbone. And the effectiveness of our method can be adequately validated.

6. Conclusion

In this work, we present a novel adaptive class suppression loss (ACSL) for long-tail datasets. ACSL is able to prevent the classifiers of tail categories from being over-suppressed by the samples from head categories. In the meantime, it preserves the discriminative power between easily confused categories. Equipped with ACSL, the models can achieve higher precision, especially on tail classes. Experimental results on long-tail datasets LVIS and Open Images validate its effectiveness. We will also apply the proposed ACSL to the long-tail image classification datasets in the future work.

Acknowledgement: This work was supported by the Research and Development Projects in the Key Areas of Guangdong Province (No.2019B010153001) and National Natural Science Foundation of China under Grants No.61772527, No.61976210, No.61807686, No.62002357, No.61806200, No.62002356, No.51975044 and No.61633002.
References


3

[26] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition, 2020.


