C2AM Loss: Chasing a Better Decision Boundary for Long-Tail Object Detection

Tong Wang[1,2], Yousong Zhu[1], Yingying Chen[1], Chaoyang Zhao[1,4], Bin Yu[1,2], Jinqiao Wang[1,2,3], Ming Tang[1]

[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] Peng Cheng Laboratory, Shenzhen, China
[4] Development Research Institute of Guangzhou Smart City, Guangzhou, China

{tong.wang,yousong.zhu,yingying.chen,chaoyang.zhao,bin.yu,jqwang,tangm}@nlpr.ia.ac.cn

Abstract

Long-tail object detection suffers from poor performance on tail categories. We reveal that the real culprit lies in the extremely imbalanced distribution of the classifier's weight norm. For conventional softmax cross-entropy loss, such imbalanced weight norm distribution yields ill conditioned decision boundary for categories which have small weight norms. To get rid of this situation, we choose to maximize the cosine similarity between the learned feature and the weight vector of target category rather than the inner-product of them. The decision boundary between any two categories is the angular bisector of their weight vectors. Whereas, the absolutely equal decision boundary is sub-optimal because it reduces the model's sensitivity to various categories. Intuitively, categories with rich data diversity should occupy a larger area in the classification space while categories with limited data diversity should occupy a slightly small space. Hence, we devise a Category-Aware Angular Margin Loss (C2AM Loss) to introduce an adaptive angular margin between any two categories. Specifically, the margin between two categories is proportional to the ratio of their classifiers' weight norms. As a result, the decision boundary is slightly pushed towards the category which has a smaller weight norm. We conduct comprehensive experiments on LVIS dataset. C2AM Loss brings 4.9∼5.2 AP improvements on different detectors and backbones compared with baseline.

1. Introduction

Object detection is one of the most essential tasks in computer vision [11, 16, 31, 32]. Modern object detectors [1, 3, 6, 12, 21, 22, 24, 31, 40, 44, 45] have achieved promising results on challenging PASCAL VOC [9] and COCO [23] datasets. Both of these two benchmarks are curated to keep the relative balance between categories. However, in real-world scenarios, data always obeys the Zipfian [29] distribution where a large number of tail categories have few samples. Although current detectors perform well on balanced datasets, they all suffer from severe performance degradation on tail classes when facing extremely imbalanced datasets. Thus, long-tail object detection remains a major challenge for researchers.

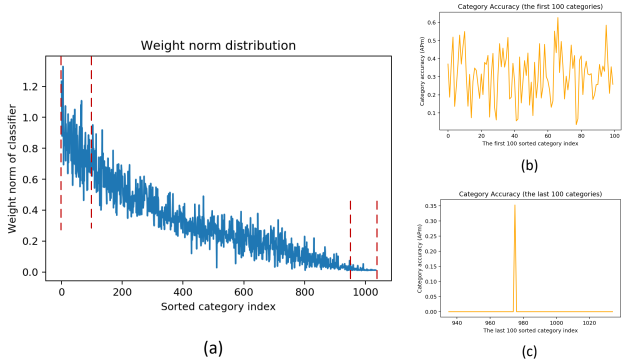


Figure 1. (a) is the classifier's weight norm distribution of a naive Mask R-CNN model trained with LVIS v1.0 training split [13]. X-axis is the sorted category index based on the category frequency. Y-axis shows the weight norm; (b) shows the precisions AP^m of the first 100 categories. (c) shows the precisions of the last 100 categories.

A model that minimizes empirical risk on long-tail training datasets is seriously biased towards head categories since they contribute most of the training data. To overcome this issue, previous literatures typically adopt two types of measures, namely, data re-sampling [8, 14, 27, 33] and loss re-weighting [34, 35, 38, 39]. Data re-sampling pins hope on adjusting the extremely imbalanced data distribution to a less imbalanced one by over-sampling the tail categories and under-sampling the head. Whereas, it only increases the occurrence frequency of tail categories. The data diversity remains unchanged, which will lead to over-fitting on tail classes. Besides, under-sampling the head classes has a risk of missing discriminative information. Loss re-weighting methods work by enhancing the loss of tail categories and weakening that of head categories. Both of these methods implicitly reshape the decision boundary and bring benefits to tail categories. Nevertheless, they adjust the decision boundary in an indirect way which may weaken their effectiveness. What's more, how they influence the decision boundary is not intuitive and geometrically interpretable.

Under the long-tail setting, we observe that the weight norm of the classifier also exhibits an extremely imbalanced distribution as shown in Fig. 1(a). This phenomenon has also been validated by previous literatures [20, 34, 35]. And we also notice the precision is highly related to the classifier's weight norm. As shown in Fig. 1(c), the weight norms of the last 100 categories are close to zero. And their precisions are almost zero. For categories that have large weight norms, their precisions vary in a reasonable range, as in Fig. 1(b). We demonstrate that the extremely imbalanced weight norm distribution will deteriorate the decision boundary, leading to a near zero precision for categories which have small weight norms. For inner-product based softmax, the output logit (before softmax) of category i is given by $||W_i||_2 \cdot ||x||_2 \cdot cos(\theta_i)$, where W_i, x, θ_i are the classifier weight, the feature and the angle between them, respectively. When $||W_i||_2$ is overwhelmingly large, the model has a high probability to predict a large score on category i. As a result, the categories with small weight norms are completely suppressed, which is fatal to their accuracy. We will detailedly analyse how the extremely imbalanced weight norm distribution causes the ill conditioned decision boundary in the following section.

The cosine classifier has natural advantages for handling the ill conditioned decision boundary mentioned above. The decision boundary of two categories is the angular bisector of the angle between two classifiers' weight vectors, as shown in Fig. 2(b). Whereas, totally abandoning the weight norm information is suboptimal since it reduces the model's sensitivity to different categories. Intuitively, categories with rich data diversity should occupy a larger area in the angular classification space. And for categories with limited data diversity, it is beneficial to slightly shrink the angular classification space for learning a compact and intrinsic feature representation. In other words, ***proper** classifier bias is profitable in long-tail object detection.*

In this paper, we propose a Category-Aware Angular Margin Loss (C2AM Loss) to adaptively adjust the decision boundary based on the weight norm distribution. Specifically, it introduces a category-aware margin to any two categories in the angular space. The angular margin is proportional to the ratio of the classifier's weight norm. We can adaptively push the decision boundary towards categories which have smaller weight norms to learn a more compact and intrinsic feature representation. Noting that although C2AM Loss manually introduces the classifier bias to the model, it will not generate ill conditioned decision boundary like the inner-product based softmax loss. C2AM Loss utilizes a hyper-parameter α to control the strength of pushing the decision boundary. Besides, a convex function $log(x)$ is utilized to ensure the margin will not become excessively large. The above two measures guarantee the classifier bias is maintained in a proper magnitude.

To validate the effectiveness of C2AM Loss, we conduct extensive experiments on the challenging long-tail object detection dataset LVIS (v0.5 and v1.0) [13]. Experimental results of various detectors (Mask R-CNN [15] and Cascade Mask R-CNN [1]) with different backbones (ResNet-50 and ResNet-101 [16]) all show the superiority of the proposed C2AM Loss. To be more specific, Mask R-50 with C2AM Loss outperforms the baseline by 5.2 AP_m. The improvements are mainly from rare categories (+11.9 AP_r^m) and common categories (+6.8 AP_c^m). We also compare our methods with other SOTA methods and the results show that our method is more competitive.

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

1. We point out that the extremely imbalanced weight norm distribution under the long-tail setting yields ill conditioned decision boundary, which severely deteriorates the performance.

2. We present a Category-Aware Angular Margin Loss (C2AM Loss) that can adaptively adjust the decision boundary for learning a more compact and intrinsic feature representation.

3. We conduct comprehensive experiments on long-tail object detection dataset LVIS (v0.5 and v1.0). C2AM Loss brings obvious performance improvement (4.9%∼5.2% AP^m) when compared with baseline and achieves new state-of-the-art on both LVIS v0.5 and v1.0.

2. Related Work

Object Detection. Recent years have witnessed rapid development in object detection area. Current popular object detectors can be divided into two types, one-stage and two-stage approaches. CNN based two-stage detectors [6,12,15,21,31] first generate coarse bounding box candidates by a lightweight Region Proposal Network (RPN). Then, the region features of these proposals are extracted through RoI Pooling or RoI Align operation. These features are further utilized for accurate classification and bounding box regression. One-stage detectors have a much concise pipeline. Typical one-stage approaches include SSD [24], YOLO [28], RetinaNet [22] and CornerNet [18] *etc*. They directly make predictions on the dense anchors or points without generating bounding box proposals. Since one-stage detectors do not extract region features for each proposal, they enjoy higher efficiency and are widely applied in real-world scenarios. These detectors perform well on balanced datasets. However, directly applying them to long-tail datasets obtains inferior performance due to the issues mentioned before. Thus, we intend to improve the detectors' performance on long-tail datasets.

Long-tail Recognition. Re-sampling strategy is a useful technique for imbalanced datasets. Repeat factor sampling [13] and class-aware sampling [33] aim to balance the data distribution by sampling the tail categories in a higher frequency. Special loss function is another technical direction for tackling the long-tail problem. LDAM [2] enforces class-dependent margins based on label frequencies and encourages tail classes to have larger margins. To protect the tail categories from being over-suppressed, EQL [35] ignores the negative gradients from head samples. The advanced EQL v2 [34] starts from the perspective of gradient balance. It introduces a novel gradient-guided re-weighting mechanism to keep the balance between positive and negative gradients for each classifier. ACSL [39] proposes to only suppress those semantically similar categories to protect the tail categories and to maintain the discriminative power of the network. In addition to these special functions, measures like decouple training [17], category grouping [20] also work well under the long-tail setting. All these methods implicitly reshape the decision boundary to protect the tail categories. Whereas, such an indirect way may weaken their effectiveness. Hence, we choose to adjust the decision boundary explicitly.

Margin-based Loss Functions Margin-based loss functions play an important role in metric learning and are widely adopted in tasks such as face verification and person Re-ID. To encourage intra-class compactness and inter-class separability, L-Softmax [26] loss incorporates a preset constant m multiplying with the angle between the feature and the ground-truth classifier. ArcFace [7] adds an additive angular margin to the target angle to obtain highly discriminative features for face recognition. CosFace [37] introduces a cosine margin term to further maximize the decision margin in the angular space. SphereFace [25] improves L-Softmax by normalizing the weights, which achieves better performance on a series of face recognition benchmarks. These loss functions introduce various margins to encourage discriminative learning. Nevertheless, the margin they utilized is a constant value that does not consider the characters of the classifiers. This is the main difference between C2AM Loss and these loss functions.

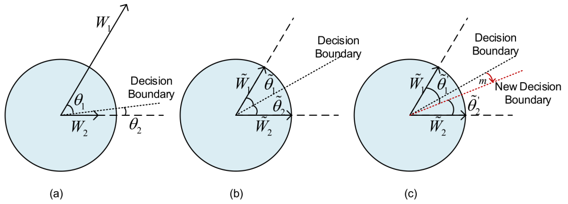


Figure 2. (a) shows the decision boundary of the conventional softmax loss; (b) illustrates the decision boundary of cosine distance based softmax loss, $\tilde{W}_1, \tilde{W}_2$ are the normalized weight vectors; (c) is the decision boundary of our proposed C2AM Loss function.

3. Methodology

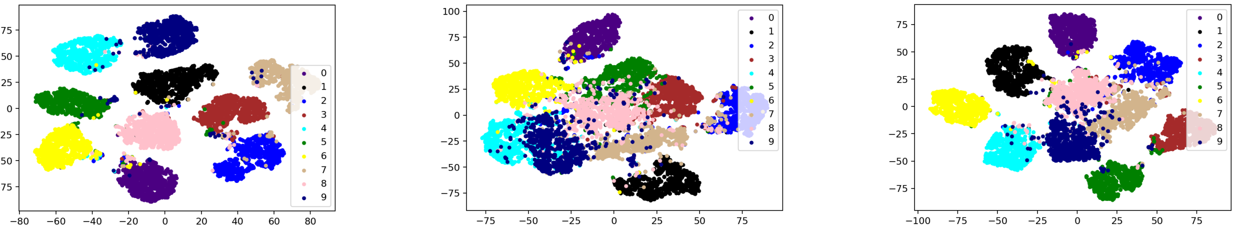
In this section, we first reveal that the extremely imbalanced weight norm distribution in long-tail recognition will generate ill conditioned decision boundary with the conventional inner-product based softmax cross-entropy loss (Sec. 3.1). And we demonstrate cosine similarity based softmax loss is helpful for getting rid of the ill conditioned decision boundary. To learn a more compact and intrinsic feature representation for tail categories, we propose a Category-Aware Angular Margin Loss (C2AM) Loss to push the decision boundary towards tail categories. Details in Sec. 3.2. To better illustrate how C2AM Loss influences the decision boundary, we perform a toy example on MNIST [19] dataset and visualize the feature distribution in Sec. 3.3. Finally, we discuss the differences between C2AM Loss and other margin-based loss functions in Sec. 3.4.

3.1. Ill Conditioned Decision Boundary of inner-product based Softmax Cross-Entropy Loss

We start by giving a review of the conventional inner-product based softmax cross-entropy loss. Given the learned feature x and ground truth i, the loss is calculated based on Eq. (1), where W_j is the j-th column of the last fully connected layer (the weight vector of classifier j). For simplicity, we omit the bias term in the last fc layer. Actually, it brings no difference to the model performance. To make a correct prediction, the model has to output the highest posterior probability of the ground-truth class, which means $W_i^T x > W_j^T x$ for all categories $j \neq i$.

$$L = -log(\frac{e^{W_i^T x}}{\sum_{j=1}^{C} e^{W_j^T x}}) \tag{1}$$

Considering the most simple binary-classification problem, the decision boundary is defined by Eq. (2). We reformulate it to Eq. (3), where θ_i is the angle between W_i and



(a) Softmax cross-entropy loss under balanced setting.

(b) Softmax cross-entropy loss under imbalanced setting.

(c) C2AM loss under imbalanced setting.

Figure 3. The feature distribution on MNIST [19] val dataset. Balanced MNIST is the original train dataset. We create an imbalanced MNIST by randomly selecting 100 images for class '7', '8', and '9'. (100 for each category, 300 in total). For classes '0'-'6', we keep all their training images. Since the feature dimension is 50, we utilize t-SNE to reduce the feature dimension to 2 for visualization.

x. Considering the situation when $||W_1||_2 \neq 0, ||W_2||_2 \neq 0, ||x||_2 \neq 0$ and $0 \leq \theta_1, \theta_2 \leq \frac{\pi}{2}$, the formulation can be further simplified to Eq. (4). As we mentioned before, long-tail datasets yield a highly imbalanced distribution of the classifier's weight norm. Supposing we have $||W_1||_2 > ||W_2||_2$, the decision boundary will move towards W_2. As a result, The angular classification space of category 2 will be shrank. As shown in Fig. 2(a), the angle (θ_2) between decision boundary and W_2 is smaller than θ_1. And things will be worse when the distribution of weight norm is extremely imbalanced. When $||W_1||_2 \gg ||W_2||_2$, θ_2 tends to be zero. For samples from category 2, the angle between the feature x and the weight W_2 must be small enough to be correctly classified. Under this circumstance, the angular classification space of category 2 will be shrank too much so that the classifier is not able to output high scores for tail samples. The tail categories are obliterated by the head.

$$W_1^T x = W_2^T x \tag{2}$$

$$||W_1||_2 \cdot ||x||_2 \cdot cos(\theta_1) = ||W_2||_2 \cdot ||x||_2 \cdot cos(\theta_2) \tag{3}$$

$$cos(\theta_1) = \frac{||W_2||_2}{||W_1||_2} cos(\theta_2) \tag{4}$$

Although the above analysis is built on the binary-class case, it is trivial to generalize the analysis to the multi-class case. In the long-tail object detection task, the imbalance factor is usually large. Head categories contain tens of thousands of instances while the instance number of tail categories is less than 100. The models trained with conventional softmax loss have ill conditioned decision boundary for tail categories. The classifier is unresponsive to the tail classes, thus yields a near-zero precision.

3.2. Category-Aware Angular Margin Loss

To get rid of the ill conditioned decision boundary under the long-tail setting, we replace the inner-product operation in the conventional softmax loss with cosine distance. Softmax loss with cosine distance minimizes the angle between feature vector x and the ground-truth classifier weight vector W_i rather than maximizing the inner-product of x and W_i. It is mathematically formulated as Eq. (5), where $cos(\theta_i) = \frac{W_i^T x}{||W_i||_2 \cdot ||x||_2}$. Here we introduce a hyper-parameter s to stabilize the training like CosFace [37] and ArcFace [7]. From this formulation, we observe that the decision boundary is only related to the angle θ, which protects the tail classifier from being over-suppressed by the head categories with extremely large weight norms. For binary-class situation, the decision boundary is the angular bisector of weight vector W_1 and W_2, as shown in Fig. 2 (b).

$$L = -log(\frac{e^{s \cdot cos(\theta_i)}}{\sum_{j=1}^{C} e^{s \cdot cos(\theta_j)}}) \tag{5}$$

Although optimizing the cosine similarity relieves the pressure of imbalanced weight norm distribution, we argue that the absolutely equal decision boundary between head and tail categories is also detrimental for the overall performance. Completely abandoning the weight norm information is irrational since it reduces the sensitivity of the model to different categories. Intuitively, head categories should occupy a larger area in the angular classification space because of the rich diversity of data. On the contrary, since the scarcity of data, the angular classification space for tail categories should be slightly shrunk to learn a more compact and intrinsic feature representation. The decision boundary should better be flexibly adjusted based on the classifier's states.

$$L_{C2AM} = -log(\frac{e^{s \cdot cos(\theta_i)}}{e^{s \cdot cos(\theta_i)} + \sum_{j=1, j \neq i}^{C} e^{s \cdot cos(\theta_j + m_{ij})}}) \tag{6}$$

where,

$$m_{ij} = max(0, \frac{\alpha}{\pi} log(\frac{||W_i||_2}{||W_j||_2})) \tag{7}$$

To this end, we reintroduce the weight norm component to the cosine classifier in a more controllable and gentle way. Specifically, we add a Category-Aware Angular Margin to the cosine similarity based softmax Loss (abbreviated as C2AM Loss). The math formulation is shown as Eq. (6). For samples from category i, C2AM Loss adds a class-aware angular margin m_{ij} to category $j (j \neq i)$, where m_{ij} is proportional to the ratio of the classifier's weight norm as Eq. (7). Noting that we detach the gradients of W_i, W_j when calculating the margin m_{ij}. We still take the binary-classification case as an example to illustrate how C2AM Loss influences the decision boundary. For samples from category 1, supposing the angle between W_1 and W_2 is t, the decision boundary of C2AM Loss is given by $cos(\theta_1) = cos(\theta_2 + m_{12})$. Since $\theta_1 + \theta_2 = t$, the decision boundary is actually $\theta_1 = \frac{t+m_{12}}{2}$. When $||W_1||_2 = ||W_2||_2$, $m_{ij} = 0$, there is no additional margin to category 2. The decision boundary is the angular bisector ($\theta_1 = \frac{t}{2}$). When $m_{12} > 0$, the decision boundary $\frac{t+m_{12}}{2}$ is larger than the angular bisector, as shown in Fig. 2 (c). The decision boundary is pushed towards the classifier weight vector with smaller weight norm. The adaptive margin m_{ij} is in proportion to the ratio of the weight norm $\frac{||W_i||_2}{||W_j||_2}$. C2AM Loss will push the decision boundary harder if the gap between the weight norms becomes larger.

It is worth noticing that although C2AM Loss pushes the decision boundary towards classifier with smaller weight norm, it will not generate ill conditioned decision boundary like the inner-product based softmax loss. First, C2AM Loss is more controllable. It introduces a hyper-parameter α to control the strength of pushing the decision boundary. α is typically a small value in our experiments. Second, it works in a more gentle way. The $log(x)$ function will output a value smaller than the input $\frac{||W_i||_2}{||W_j||_2}$. Besides, since the second derivative of $log(x)$ is smaller than 0, the output will increase slower as the input becomes larger. Overall, the above two reasons guarantee that C2AM Loss will not generate ill conditioned decision boundary.

3.3. Visualization of Toy Example

To investigate how the imbalanced data distribution influences the feature learning and validate the effectiveness of C2AM Loss, we conduct a toy example on MNIST and visualize the feature distribution in Fig. 3. For better visualization, we reduce the feature dimension from 50 to 2 with t-SNE. We first train the network on the balanced MNIST train and visualize the feature distribution of the val dataset. As shown in Fig. 3(a), although there are some false positives, we can still observe the clear decision boundaries between different classes. To illustrate how the imbalanced data distribution influences the feature distribution, we create an imbalanced MNIST train dataset by manually reducing the image number of '7', '8', and '9' to 100. As illustrated in Fig. 3(b), the decision boundary between tail category and head category becomes blurry. The feature points near the decision boundary are not discriminative, leading to many false positives. Comparing Fig. 3(b) and Fig. 3(c), we observe a clearer decision boundary in Fig. 3(c) and the features of tail categories are more discriminative in 2-dimension feature space. The above observations prove that C2AM Loss is able to encourage the model to learn a more discriminative and intrinsic feature representation.

Table 1. Comparison with other margin-based loss functions.

Loss Function	Formulation
CosFace [37]	$L = -log(\frac{e^{s \cdot (cos(\theta_i) - m)}}{e^{s \cdot (cos(\theta_i) - m)} + \sum_{j=1, j \neq i}^{C} e^{s \cdot cos(\theta_j)}})$
ArcFace [7]	$L = -log(\frac{e^{s \cdot cos(\theta_i + m)}}{e^{s \cdot cos(\theta_i + m)} + \sum_{j=1, j \neq i}^{C} e^{s \cdot cos(\theta_j)}})$
SphereFace [25]	$L = -log(\frac{e^{\|\|x_i\|\| \psi(\theta_i)}}{e^{\|\|x_i\|\| \psi(\theta_i)} + \sum_{j=1, j \neq i}^{C} e^{\|\|x_i\|\| cos(\theta_j)}})$
$\psi(\theta_i) = (-1)^k cos(m\theta_i) - 2k, \theta_i \in [\frac{k\pi}{m}, \frac{(k+1)\pi}{m}], k \in [0, m-1]$	
C2AM Loss	$L = -log(\frac{e^{s \cdot cos(\theta_i)}}{e^{s \cdot cos(\theta_i)} + \sum_{j=1, j \neq i}^{C} e^{s \cdot cos(\theta_j + m_{ij})}})$
	$m_{ij} = max(0, \frac{\alpha}{\pi} log(\frac{\|\|W_i\|\|_2}{\|\|W_j\|\|_2}))$

3.4. Discussion

Although C2AM Loss shares a similar formulation with other margin-based loss functions, they are designed with totally different motivations. CosFace [37], ArcFace [7] and SphereFace [25] introduce a preset margin m to maximize inter-class variance and minimize intra-class variance. However, C2AM Loss designs an adaptive margin to adjust the decision boundary between head and tail categories. To better distinguish our method from others, we list their math formulations in Table 1. For CosFace and ArcFace, the margin is introduced in an additive manner. CosFace adds a negative preset margin to the cosine similarity $cos(\theta_i)$. While ArcFace directly adds the constant margin m to the angular θ_i. In addition to the additive manner, the constant margin can also be multiplied to the angular θ_i, as the SphereFace does. Formally speaking, our proposed C2AM Loss looks more like ArcFace which all add an additional

Table 2. Performance comparison of Cross-Entropy Loss and C2AM Loss on LVIS v1.0 `val`.

Framework	Backbone	Loss	AP^m	AP^b	AP_r^m	AP_c^m	AP_f^m
Mask R-CNN	ResNet-50	Cross-Entropy	20.5	21.4	1.1	18.6	31
		C2AM Loss	**25.7**	**26.5**	**13**	**25.4**	**31.5**
	ResNet-101	Cross-Entropy	21.8	22.8	1.4	20.3	32.5
		C2AM Loss	**27**	**28.1**	**14.1**	**26.7**	**33**
Cascade Mask R-CNN	ResNet-50	Cross-Entropy	22.7	25.3	2.8	21.6	32.7
		C2AM Loss	**27.6**	**31.1**	**14.2**	**27.7**	**33.5**
	ResNet-101	Cross-Entropy	24.3	27	3.3	23.7	34.1
		C2AM Loss	**29.2**	**32.6**	**16.8**	**29.1**	**34.7**

margin to the angular. However, the essential difference lies in that the margin in C2AM Loss is **adaptive**, which is reflected in the following two aspects:

First, the margin in C2AM Loss is category-aware. For CosFace, ArcFace, and SphereFace, the margins between any two categories are the same value m. However, the margin in C2AM Loss is related to the classifier's weight norm, which yields various margins between different categories. *Second, the margin in C2AM Loss will change as the training goes on.* During training, the network's parameters will be updated. The dynamically changing classifier yields an adaptive angular margin m, which is beneficial for precisely adjusting the decision boundary.

4. Experiments

4.1. Dataset and Evaluation Metric

To validate the effectiveness of our proposed C2AM Loss, we conduct comprehensive experiments on the long-tail Large Vocabulary Instance Segmentation (LVIS) dataset [13]. LVIS provides precise bounding box and mask annotations for various categories with long-tail distribution. We mainly perform experiments on the v1.0 version which consists of 1203 categories. The whole dataset is split to `train` set (100k images with 1.3M instances) and `val` set (19.8k images). We train our models on `train` set and report the accuracy on `val` set. LVIS divides all categories into 3 groups based on the their frequency in the `train` set: rare (<10 images), common (11−100 images) and frequent (>100 images). For evaluation, we report the mean average precision (AP^m for mask prediction, AP^b for box prediction). Besides, the average precision on rare (AP_r^m), common (AP_c^m) and frequent (AP_f^m) categories are also reported to well characterize the long-tail class performance. In addition to LVIS v1.0, we also release the results on LVIS v0.5 for comparison.

4.2. Implementation Details

We implement our methods with the popular MMDetection [4] toolbox and mainly conduct experiments on Mask R-CNN [15] detector. ResNet50 [16] with FPN [21] architecture has been adopted as the backbone. Besides, we also perform experiments with a larger backbone network, such as ResNet101, to validate the effectiveness of the C2AM Loss. When training, we choose end-to-end training with 2x training schedule. The models are trained using SGD optimizer with 0.9 momentum and 0.0001 weight decay and batch size of 16 on 8 GPUs. The initial learning rate is set to 0.02 with 500 iterations' warm up. The learning rate decays to 0.002, 0.0002 at the end of epoch 16 and 22, respectively. The training stops at epoch 24. Following the convention, we apply random horizontal image flipping and multi-scale jittering with smaller image sizes (640, 672, 704, 736, 768, 800) in all experiments. When testing, the image size is set to (1333, 800) without any test time augmentation. Non-Maximum Suppression is performed with IoU threshold 0.5 to remove duplicates. After NMS, the top 300 bounding boxes with score threshold 0.0001 for per image are selected for evaluation. When combining C2AM Loss with Mask R-CNN, we simply replace the cross-entropy loss on top of the bounding box classification branch with C2AM Loss.

4.3. Main Results

To validate the effectiveness of C2AM Loss, we conduct experiments with Mask R-CNN and Cascade Mask R-CNN [1] with various backbones, ResNet-50 and ResNet-101. We train the baseline model with cross-entropy loss for 24 epochs. The experimental results are summarized in Table 2. The baseline model (Mask R-50) has a rather imbalanced accuracy distribution. Frequent categories have satisfactory precision (31%) while the accuracy of rare categories is almost zero (1.1%). The extremely imbalanced weight norm distribution severely deteriorates the decision boundary of tail categories. Hence, the model is not able to correctly classify the samples of tail classes. With C2AM Loss, the precision of tail categories AP_r^m is greatly improved by a large margin (+11.9%). Besides, we can also observe obvious performance improvement for AP_c^m (+6.8%), which is consistent with our analysis. What's

Table 3. Results of C2AM Loss with different hyper-parameter s on LVIS v1.0 `val`. The model is Mask R-CNN with ResNet-50 backbone.

s	AP^m	AP^b	AP_r^m	AP_c^m	AP_f^m
10	14.2	14.9	0	7.6	27.7
20	25.4	26.1	12.5	25	31.6
30	**25.7**	**26.5**	13	25.4	31.5
40	24.8	25.5	13.2	23.9	30.8
50	23.6	24.4	11.5	22.7	30

Table 4. Results of C2AM Loss with different hyper-parameter α on LVIS v1.0 `val`. The model is Mask R-CNN with ResNet-50 backbone.

α	AP^m	AP^b	AP_r^m	AP_c^m	AP_f^m
0.0	24.1	24.8	9.6	23.2	31.5
0.1	24.8	25.5	11.6	24.2	31.4
0.3	25.3	26.1	12.8	24.7	31.6
0.5	**25.7**	**26.5**	13	25.4	31.5
0.7	25.4	26.4	13	24.9	31.5

Table 5. Results of Mask-R-50 with different margin type on LVIS v1.0 `val`.

Margin Type	AP^m	AP^b	AP_r^m	AP_c^m	AP_f^m
None	24.1	24.8	9.6	23.2	31.5
Adaptive	**25.7**	**26.5**	**13**	**25.4**	**31.5**
Fixed	24.3	25.2	11.5	23.4	30.9

more, C2AM Loss improves the tail categories without sacrificing the head. Actually, AP_f^m has little improvement from 31% to 31.5%. The overall precision AP^m is lifted by a significant margin (+5.2%).

When switching to a large model, C2AM Loss still brings consistent performance improvements. For Mask R-CNN with ResNet-101 backbone, the overall accuracy for mask prediction AP_m reaches 27% and AP_b increases by 5.3%, from 22.8% to 28.1%. The precision increases of AP_r^m (12.7%) and AP_c^m (6.4%) are still significant. It is worth noticing that C2AM Loss does not hurt the performance of head categories, which is a desired property for long-tail solutions. In principle, the effectiveness of C2AM Loss is not restricted to a certain type of detector. To verify that, we conduct experiments with the more powerful Cascade Mask R-CNN detector. By simply replacing the original softmax cross-entropy loss on all 3 heads with our proposed C2AM Loss, the performance can be greatly boosted by a large margin, especially for tail categories. The details are in Table 2. And the results demonstrate that C2AM Loss is versatile for different object detectors.

4.4. Ablation Study

Ablation Study on Hyper-parameters. C2AM Loss introduces two hyper-parameters s and α. s is a scale factor for efficiently optimizing the cosine similarity based softmax loss. It is a standard configuration for the cosine classifier and has been widely used in various verification tasks such as face recognition and person re-identification. NormFace [36] demonstrated that cosine similarity with softmax loss is hard to optimize since the range of cosine value is limited, [-1,1]. The low range problem may prevent the probability $P_i = \frac{e^{cos\theta_i}}{\sum_{j=1}^{C} e^{cos\theta_j}}$ from getting close to 1 even when the samples are well-separated. Introducing s to scale the cosine value to a proper magnitude is necessary for stable optimization. We carefully tune this hyper-parameter and record the results in Table 3. We found the best setting is 30, which is consistent with the recommended setting of CosFace [37], ArcFace [7] *et al.*

Another hyper-parameter of C2AM Loss is α. It controls the strength of how hard we push the decision boundary. If α is set too small, the strength is too weak to influence the final decision boundary. When α is set to 0, C2AM loss degenerates to the cosine classifier combining with softmax cross-entropy loss. We conduct experiments with different α and list the results in Table 4. We observe that C2AM Loss outperforms the cosine classifier with an obvious precision rise (+1.6% AP^m, +1.7% AP^b). We experimentally find that $\alpha = 0.5$ works best. So we adopt this default setting to conduct all experiments related to C2AM Loss.

Adaptive Margin or Fixed Margin? Since C2AM Loss sets adaptive margin between categories, a natural question will be: what will happen if we set a fixed margin just like CosFace and ArcFace? In order to illustrate the necessity of the adaptive margin, we design control experiments about the type of margin, namely, adaptive margin and fixed margin. For fixed margin, we replace the adaptive margin term $\frac{\alpha}{\pi}log(\frac{||W_i||}{||W_j||})$ with a constant value m. After elaborately tuning the value of m, we find the fixed margin works worse than adaptive margin, Table 5. It obtains worse performance on both AP_r^m, AP_c^m and AP_f^m, which indicates that the fixed margin can not effectively adjust the decision boundary. Since the fixed margin ignores the characters of different categories, it is not suitable for all classes. Thus, the category-aware margin is necessary under the long-tail setting.

4.5. Comparison with State-of-the-Arts

In this section, we compare our method with other state-of-the-art methods, as shown in Table 6. Since LVIS v1.0 is a newly released dataset, we also report the results on LVIS v0.5 for comparison with more previous methods. Our models are trained with repeat factor sampler for 24 epochs. There is no test time augmentation during testing. For LVIS v0.5, we present the results of Mask R-CNN with ResNet50-FPN backbone. C2AM Loss performs better than other methods, outperforming the state-of-the-art method

Table 6. Comparison with state-of-the-art methods on LVIS v0.5 and LVIS v1.0 dataset. **Bold** numbers denote the best results.

Dataset	Framework	Backbone	Methods	AP^m	AP^b	AP_r^m	AP_c^m	AP_f^m
LVIS v0.5	Mask R-CNN	R-50-FPN	CBL [5]	23.3	23.9	11.4	23.8	27.3
			LWS [17]	23.8	24.1	14.4	24.4	26.8
			LDAM [2]	24.1	24.5	14.6	25.3	26.3
			EQL [35]	25.2	24.1	14.6	24.4	26.8
			Forest R-CNN [41]	25.6	25.9	18.3	26.4	27.6
			RFS [13]	25.9	26.1	17.8	26.2	28.8
			BAGS [20]	26.3	25.8	18.0	26.9	28.7
			BALMS [30]	27.0	27.6	19.6	28.9	27.5
			EQLv2 [34]	27.1	27.0	18.6	27.6	29.9
			DisAlign [43]	27.9	27.6	16.2	29.3	30.8
			LOCE [10]	28.4	28.2	**22.0**	29.0	30.2
			C2AM Loss (Ours)	**29.7**	**29.8**	19.3	**31.3**	**31.8**
LVIS v1.0	Mask R-CNN	R-50-FPN	RFS [13]	23.7	24.7	13.5	22.8	29.3
			FASA [42]	24.4	24.2	15.4	23.5	29.4
			EQLv2 [34]	25.5	26.1	17.7	24.3	30.2
			Seesaw Loss [38]	26.4	27.4	**19.6**	26.1	29.8
			LOCE [10]	26.6	27.4	18.5	26.2	30.7
			C2AM Loss (Ours)	**27.2**	**27.9**	16.6	**27.2**	**31.9**
	Mask R-CNN	R-101-FPN	RFS [13]	25.5	26.6	16.6	24.5	30.6
			FASA [42]	26.3	27.0	19.1	25.4	30.6
			EQLv2 [34]	27.2	27.9	**20.6**	25.9	31.4
			LOCE [10]	28.0	29.0	19.5	27.8	32.0
			Seesaw Loss [38]	28.1	28.9	20.0	28.0	31.9
			C2AM Loss (Ours)	**28.6**	**29.4**	18.1	**28.5**	**33.2**

LOCE by 1.3% AP^m, 1.6% AP^b. We notice that C2AM Loss achieves the highest precision on AP_c^m and AP_f^m. We attribute it to the ability of C2AM Loss to adaptively adjust the decision boundary between the head and tail categories. For LVIS v1.0, we present the results of Mask R-CNN with ResNet50 and ResNet101 backbone. For Mask-R-50, C2AM Loss still obtains the best performance with 27.2%AP_m and 27.9%AP_b, suppressing other methods, including EQLv2, LOCE and Seesaw Loss. Similarly, the advantages of C2AM Loss on head categories are obvious, 1.2% AP_f^m higher than LOCE. In terms of larger backbone ResNet101, C2AM Loss can also achieve the best performance 28.6% AP^m, suppressing the current sota method Seesaw Loss by 0.5% AP^m and AP^b. Although C2AM Loss doesn't achieve the best result of AP_r^m, it obtains the best result on both AP_c^m and AP_f^m, leading to the highest overall performance AP^m and AP^b. We conjecture the reason lies in that other methods pursue the highest performance on tail categories at a cost of sacrificing the head. While our method focuses on both the head and the tail, thus can achieve the best overall performance AP^m and AP^b.

5. Limitations

C2AM Loss shares a similar math formulation with other margin-based loss functions, CosFace [37], ArcFace [7] and SphereFace [25] *et al*. However, in this paper, we mainly focus on the long-tail object detection and conduct experiments on LVIS dataset only. How does the C2AM Loss behave on other tasks, like face recognition and person re-identification, still remains unknown. This is our main limitation which needs more effort to explore.

6. Conclusion

In this paper, we reveal that the extremely imbalanced distribution of the classifier's weight norm yields an ill conditioned decision boundary for classifiers with small weight norms, thus leading to poor performance of these categories. To chase a better decision boundary for long-tail object detection, we present a category-aware angular margin loss (C2AM Loss) to adaptively adjust the decision boundary based on the classifier's weight norm. We conduct extensive experiments on the challenging LVIS dataset. The results show that C2AM Loss achieves consistent gains on various detectors and backbones. Moreover, C2AM Loss sets new state-of-the-art on both LVIS v0.5 and v1.0.

Acknowledgement: This work was supported by Key-Area Research and Development Program of Guangdong Province(No.2021B0101410003), National Natural Science Foundation of China under Grants No.62002357, No.62176254, No.61976210, No.61876086, No.62076235, No.62006230 and No.51975044, Open Research Projects of Zhejiang Lab (No.2021KH0AB07).

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, 2018. 1, 2, 6

[2] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, volume 32, pages 1567–1578. Curran Associates, Inc., 2019. 3, 8

[3] Kean Chen, Jianguo Li, Weiyao Lin, John See, Ji Wang, Lingyu Duan, Zhibo Chen, Changwei He, and Junni Zou. Towards accurate one-stage object detection with ap-loss. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5119–5127, 2019. 1

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6

[5] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019. 8

[6] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 379–387, 2016. 1, 2

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3, 4, 5, 7, 8

[8] Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, pages 1–8. Citeseer, 2003. 1

[9] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 1

[10] Chengjian Feng, Yujie Zhong, and Weilin Huang. Exploring classification equilibrium in long-tailed object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3417–3426, October 2021. 8

[11] Y. Gao, Y. Chen, J. Wang, and H. Lu. Progressive rectification network for irregular text recognition. *Sciece China. Information Sciences*, 63(2), 2020. 1

[12] Ross Girshick. Fast r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 1, 2

[13] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019. 1, 2, 3, 6, 8

[14] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *ICIC'05 Proceedings of the 2005 international conference on Advances in Intelligent Computing - Volume Part I*, pages 878–887, 2005. 1

[15] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 6

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2, 6

[17] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition, 2020. 3, 8

[18] H. Law and J. Deng. Cornernet: Detecting objects as paired keypoints. In *International Journal of Computer Vision*, pages 642–656, 2020. 3

[19] Yann LeCun, Corinna Cortes, and Christopher Burges. The mnist database of handwritten digits. 1998. 3, 4

[20] Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, and Jiashi Feng. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 3, 8

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1, 2, 6

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. 1, 3

[23] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014. 1

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. *european conference on computer vision*, pages 21–37, 2016. 1, 3

[25] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3, 5, 8

[26] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. *JMLR.org*, 2016. 3

[27] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly

supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018. 1

[28] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[29] W. J. Reed. The pareto, zipf and other power laws. *Economics Letters*, 74(1):15–19, 2001. 1

[30] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems*, 2020. 8

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. 1, 2

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael and Bernstein. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1

[33] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*, pages 467–482. Springer, 2016. 1, 3

[34] Jingru Tan, Xin Lu, Gang Zhang, Changqing Yin, and Quanquan Li. Equalization loss v2: A new gradient balance approach for long-tailed object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1685–1694, June 2021. 1, 2, 3, 8

[35] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 3, 8

[36] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: l_2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 10411049, New York, NY, USA, 2017. Association for Computing Machinery. 7

[37] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3, 4, 5, 7, 8

[38] Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw loss for long-tailed instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9695–9704, June 2021. 1, 8

[39] Tong Wang, Yousong Zhu, Chaoyang Zhao, Wei Zeng, Jinqiao Wang, and Ming Tang. Adaptive class suppression loss for long-tail object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3103–3112, June 2021. 1, 3

[40] Tong Wang, Yousong Zhu, Chaoyang Zhao, Wei Zeng, Yaowei Wang, Jinqiao Wang, and Ming Tang. Large batch optimization for object detection: Training coco in 12minutes. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 481–496, Cham, 2020. Springer International Publishing. 1

[41] Jialian Wu, Liangchen Song, Tiancai Wang, Qian Zhang, and Junsong Yuan. Forest r-cnn: Large-vocabulary long-tailed object detection and instance segmentation. *ACM*, 2020. 8

[42] Yuhang Zang, Chen Huang, and Chen Change Loy. Fasa: Feature augmentation and sampling adaptation for long-tailed instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3457–3466, October 2021. 8

[43] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2361–2370, June 2021. 8

[44] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Single-shot refinement neural network for object detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4203–4212, 2018. 1

[45] Xiaosong Zhang, Fang Wan, Chang Liu, Rongrong Ji, and Qixiang Ye. Freeanchor: Learning to match anchors for visual object detection. In *Advances in Neural Information Processing Systems*, pages 147–155, 2019. 1