IOU-AWARE MULTI-EXPERT CASCADE NETWORK VIA DYNAMIC ENSEMBLE FOR LONG-TAILED OBJECT DETECTION

Wan-Cyuan Fan[1,2∗], *Cheng-Yao Hong*[1∗], *Yen-Chi Hsu*[1,3∗], *and Tyng-Luh Liu*[1]

[1]Institute of Information Science, Academia Sinica
[2]Graduate Institute of Communication Engineering, National Taiwan University
[3]Graduate Institute of Computer Science & Information Engineering, National Taiwan University

ABSTRACT

Object detection over a long-tailed large-scale dataset is practical, challenging, and comprehensively under-explored. Recently proposed methods mainly focus on eliminating the imbalanced classification problem. However, only a few attempts have been made to consider the quality of the predicted bounding boxes. Inspired by the observation of existing Cascade architecture, "**detectors with specific IoU thresholds excel at different label frequencies of bounding boxes**," this paper first pinpoints the issue in long-tailed distribution. A detector may predict inaccurate bounding boxes on the categories of fewer training data such that the corresponding extracted visual features could further degrade the classification accuracy. Thus, the predicted accuracy of bounding boxes becomes substantially different among categories in the long-tailed distribution. We introduce a *Multi-Expert Cascade* (MEC) framework that readjusts the weight of each category in the training process via a multi-expert loss. Furthermore, we leverage dynamic ensemble mechanisms at inference time to fully utilize expert detectors and achieve better performance. Extensive experiments on the recent long-tailed large vocabulary object detection dataset show that the proposed MEC framework significantly improves the performance of most widely-used detectors over various backbones on object detection and instance segmentation tasks.

***Index Terms*—** Object Detection, long-tailed distribution, representation learning

1. INTRODUCTION

Convolutional neural networks have enabled significant advancements in object detection [1, 2]. Object detection techniques can be categorized into two groups based on their model characteristics. One-stage detectors are known for their real-time efficiency [3, 4, 5, 6], while the state-of-the-art approach for detection involves a two-stage process of region proposal and classification. The R-CNN series [7, 8, 9, 10] has shown promising results on object detection. Popular frameworks such as [11, 12] have further improved performance by incorporating multiple detection heads with proposed region refinement.

The aforementioned methods yield promising results on manually balanced benchmarks such as COCO [13] and PASCAL VOC [14]. However, the practical data distribution tends to be highly imbalanced in the real world, with an extremely long-tailed class distribution. Under this circumstance, many existing architectures may fail to achieve the expected performance. Several proposed approaches have tried to address the imbalanced classification problem in recent years, such as [15, 16, 17, 18, 19, 20, 21, 22, 23]. In the early days, [15] introduces a dynamic sampling factor that increases the probability of sampling images with rare data during training. After that, [24, 16, 19] point out that the image-wise resampling methods will damage the representation, and then they decouple the learning procedure into representation learning and classification learning. Recently, [20] argues that the imbalanced classification problem comes from lousy momentum causal effect and proposes de-confounded training and total direct effect inference to eliminate the causal effect. Although many efforts have been made to resolve the imbalanced classification problem, the imbalanced detection problem still needs to be further explored.

The motivation is inspired by an attractive observation of Cascade architectures [11, 12], perennial winners on the leaderboard. As shown in Figure 1(a), we train various detectors with different foreground thresholds on a long-tailed dataset. The result shows that the performance of each category is correlated to the IoU threshold. For example, rare (categories with fewer training samples) instances perform better when the detector is trained with a lower IoU threshold, and vice versa. Furthermore, at the inference stage, we argue that it is hard for a detector to predict the bounding boxes with high accuracy for rare instances, which induces the detector trained with a lower IoU threshold to achieve better performance for rare categories. In short, there is an imbalanced detection problem when the Cascade architectures face the long-tailed distribution. Therefore, this paper proposes a multi-expert and IoU-aware detection framework

∗ These three authors contributed equally.

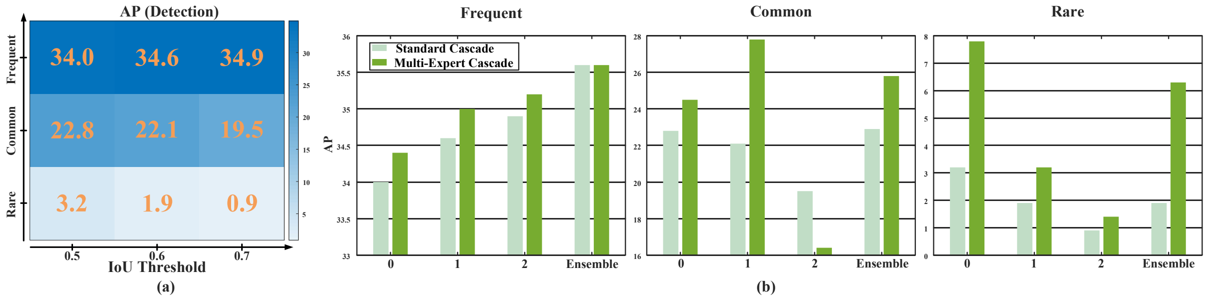

Fig. 1. (a) shows that the performance of standard Cascade architecture detector of different frequencies on different stages. Besides, it points out that there is an imbalanced detection problem in the multi-stage detector. (b) presents the performance comparison between standard Cascade and the proposed Multi-Expert Cascade approach (Both with ResNet101 as backbone).

called **M**ulti-**E**xpert **C**ascade (MEC) framework to address the above issue. First, we divide the classifier into multi-group according to the number of categories in the dataset. After that, we train the multi-stage detector with the multi-expert loss, which re-weights the gradient in the training process for each category at different stages. So that each classifier can emphasize a specific group of categories. Second, we introduce a dynamic ensemble mechanism to control the ensemble weights between these expert classifiers in the inference time to improve the effectiveness of the "expert" detectors. A comparison between the standard Cascade R-CNN and Multi-Expert Cascade on LVIS is provided in Figure 1(b). Our contributions can be summarized as follows:

- We identify the imbalanced detection problem, where different categories have varying frequencies and are sensitive at different stages in a multi-stage detector architecture.
- We present Multi-Expert Cascade (MEC), a novel end-to-end framework that consists of the *multi-expert loss* for training and the *dynamic ensemble mechanism* at inference to tackle the imbalanced detection problem and achieve better performance.

2. MULTI-EXPERT CASCADE

2.1. Multi-expert Loss

As depicted in Figure 1(a), when the distribution of categories is imbalanced, the performance of the Cascade detector on categories of different frequencies is highly dependent on the quality of the predicted bounding boxes. To leverage this characteristic, we propose the Multi-Expert Cascade (MEC) framework, which boosts the positive and negative gradient for specific categories to enable each stage to specialize in detecting particular categories. The architecture of the proposed framework is shown in Figure 2. For a K-stage architecture model H, we define H^k to represent the kth stage classifier. Next, we group all the categories into N groups based on the number of their training instances. Category c is assigned to $\mathcal{G}n$ (nth group) if $l_n \leq \alpha(i) < ln+1$, where $\alpha(i)$ denotes the number of instances of category i in the training set, and l_n and l_{n+1} are the thresholds used to determine the corresponding group. Following the setting in LVIS [15], we set $l_1 = 0$, $l_2 = 10^1$, $l_3 = 10^2$, and $l_4 = +\infty$. To enable each stage to focus on specific categories, we propose to enforce each classifier to concentrate on the group it is good at handling. Specifically, we make an additional prediction from each stage. Given an input instance feature $x \in \mathbb{R}^d$ and the corresponding one-hot label $y \in \{0,1\}^C$, we obtain the output logits $o^k \in \mathbb{R}^C, k = 1, ..., K$ from each stage classifier. Here, d represents the feature size, and C denotes the number of categories. We further define the standard prediction p_i^k and the expert prediction $\tilde{p}_i^k$ for the ith category in the kth stage group as follows:

$$\hat{p}_i^k = \begin{cases} p_i^k = \frac{e^{o_i^k}}{\sum_{j=1}^{C} e^{o_j^k}} & , i \notin \mathcal{G}_k \\ \underbrace{(\tilde{p}_i^k = \frac{e^{o_i^k}}{\sum_{j \in \mathcal{G}_k} e^{o_j^k}})^{1-\lambda}}_{\text{Expert Prediction}} \times \underbrace{(p_i^k = \frac{e^{o_i^k}}{\sum_{j=1}^{C} e^{o_j^k}})^{\lambda}}_{\text{Standard Prediction}} & , i \in \mathcal{G}_k \end{cases} , \quad (1)$$

where hyper-parameter λ is used to control the ratio between p_i^k and $\tilde{p}_i^k$. Then we define the multi-expert loss as the cross-entropy between joint prediction $\hat{p}_i^k$ and label y_i:

$$\mathcal{L}_{\text{exp}} = -\sum_{k=1}^{N} \sum_{i=1}^{C} y_i \log \hat{p}_i^k. \quad (2)$$

Through the equation (1), we can virtually amplify the performance of each head on the corresponding group by the joint probability. The whole training process mentioned above is shown in Figure 2(a).

2.2. Dynamic Ensemble Mechanism

In line with standard multi-stage methods [11, 12], the most straightforward approach to handle multi-stage predictions is to use the average ensemble mechanism. However, this approach fails to fully leverage the expertise of each stage. To

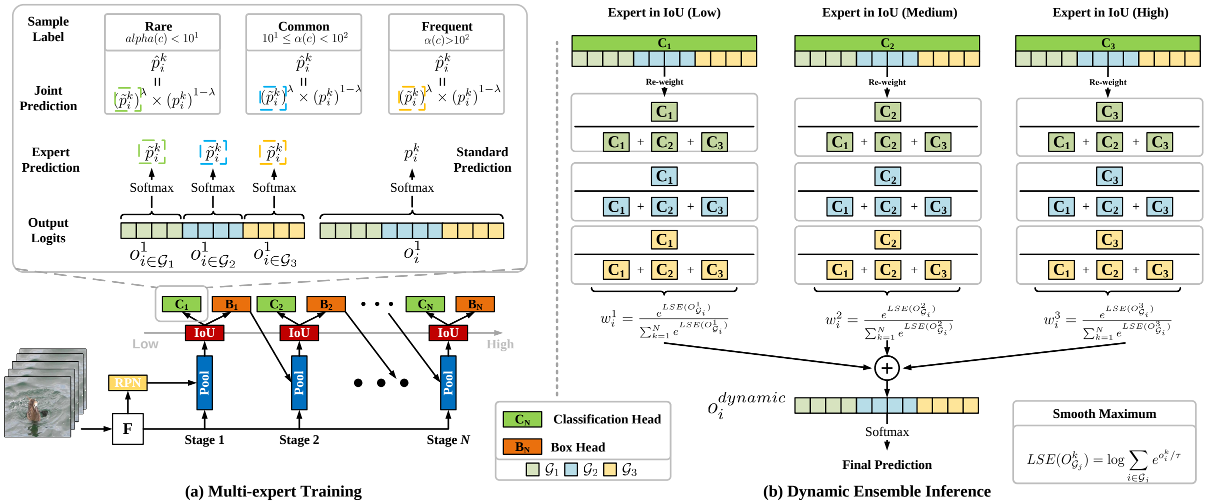
Sample Label
Rare
$alpha(c) < 10^1$
Common
$10^1 \le \alpha(c) < 10^2$
Frequent
$\alpha(c)>10^2$
Joint Prediction
$\hat{p}_i^k = (\tilde{p}_i^k)^\lambda \times (p_i^k)^{1-\lambda}$
Expert Prediction
$\tilde{p}_i^k$
Softmax
Standard Prediction
p_i^k
Output Logits
$o^1_{i\in\mathcal{G}_1}$
$o^1_{i\in\mathcal{G}_2}$
$o^1_{i\in\mathcal{G}_3}$
o^1_i
C1
B1
C2
B2
CN
BN
IoU
Low
High
RPN
Pool
F
Stage 1
Stage 2
Stage N
(a) Multi-expert Training
Expert in IoU (Low)
Expert in IoU (Medium)
Expert in IoU (High)
Re-weight
C1 + C2 + C3
$w_i^1 = \frac{e^{LSE(O^1_{\mathcal{G}_i})}}{\sum_{k=1}^N e^{LSE(O^1_{\mathcal{G}_i})}}$
$w_i^2 = \frac{e^{LSE(O^2_{\mathcal{G}_i})}}{\sum_{k=1}^N e^{LSE(O^2_{\mathcal{G}_i})}}$
$w_i^3 = \frac{e^{LSE(O^3_{\mathcal{G}_i})}}{\sum_{k=1}^N e^{LSE(O^3_{\mathcal{G}_i})}}$
$o_i^{dynamic}$
Softmax
Final Prediction
CN Classification Head
BN Box Head
$\mathcal{G}_1$ $\mathcal{G}_2$ $\mathcal{G}_3$
Smooth Maximum
$LSE(O^k_{\mathcal{G}_j}) = \log \sum_{i\in\mathcal{G}_i} e^{o_i^k/\tau}$
(b) Dynamic Ensemble Inference

Fig. 2. MEC framework consists of two components. (a) In the training phase, we leverage the observation that different orders of magnitude of categories favor different IoU thresholds to optimize each detector as the *multi-expert loss*. (b) In the inference process, we utilize an *dynamic ensemble mechanism* to effectively exert the advantages with *multi-expert loss* training.

address this issue, we propose two ensemble mechanisms that capitalize on the multi-stage setup and outperform the average ensemble mechanism. These two approaches are the sparse ensemble mechanism and the dynamic ensemble mechanism.

To begin, the sparse ensemble mechanism is based on the observation presented in Figure 1. The mechanism operates under the assumption that each stage functions as an expert for its assigned group under the multi-expert loss. As such, the inference output logit o_i^{sparse} is constructed by sparsely combining the outputs of three distinct experts, as follows:

$$o_i^{\text{sparse}} = \sum_{k=1}^{K} e_i^k o_i^k, \text{ where } e_i^k = \begin{cases} 0 & , i \notin \mathcal{G}_k \\ 1 & , i \in \mathcal{G}_k \end{cases} . \quad (3)$$

Secondly, the proposed dynamic ensemble mechanism aims to effectively utilize the unique characteristics of our Multi-expert Cascade model. Specifically, the multi-stage module is now capable of better distinguishing the input instance's category within each expert group by amplifying the difference between the highest logit and the mean of all logits. As illustrated in Figure 2(b), we first employ the LogSumExp (LSE) as the smooth maximum function to obtain the highest logit in each group. Then, the logit of the ith category on the kth stage, with the corresponding group $\mathcal{G}_j$, is computed as follows:

$$o_i^{\text{dynamic}} = \sum_{k=1}^{N} w_i^k o_i^k, \quad (4)$$

where

$$w_i^k = \frac{e^{LSE(O^k_{\mathcal{G}_j})}}{\sum_{k=1}^{N} e^{LSE(O^k_{\mathcal{G}_j})}}, i \in \mathcal{G}_j \quad (5)$$

and

$$LSE(O^k_{\mathcal{G}_j}) = \log \sum_{i\in\mathcal{G}_j} e^{o_i^k/\tau} \quad (6)$$

with temperature τ. Unlike the sparse ensemble mechanism, the dynamic ensemble mechanism takes into account the logits from each stage and represents a balance point between the average and sparse ensemble mechanisms. Following the ensemble mechanism, the output probability vector can be expressed as $S^{\mathcal{E}} = \sigma(o^{\mathcal{E}})$, where σ denotes the softmax activation function, and $\mathcal{E}$ refers to the ensemble mechanisms proposed in this study.

3. EXPERIMENTAL RESULTS AND ANALYSIS

We conduct comprehensive experiments on two datasets: the Large Vocabulary Instance Segmentation (LVIS) dataset [15] and **COCO-LT** [25] dataset. Firstly, **LVIS v1.0** comprises 100,170 training images and 19,809 validation images with 1,203 categories. These categories are divided into three groups based on the number of images that contain those categories: rare (1-10), common (11-100), and frequent (>100 images). We adopt the mean Average Precision (mAP) as the evaluation metric and report AP$_r$, AP$_c$, and AP$_f$ with the corresponding group. To further validate the MEC on long-tailed distribution, we create the **COCO-LT** dataset, a subset of COCO, following the construction methods in BAGS [25]. COCO-LT follows a long-tail distribution like LVIS and contains 16,966 training images of 80 categories, including 128,615 training instances. For the following experiments, we utilize Cascade R-CNN with a feature pyramid ResNet-50 as the default backbone unless mentioned otherwise.

Table 1. Compare the performance of the proposed MEC framework with other long-tailed approaches on LVIS v1.0.

ID	Models	$m\mathbf{AP}^m$	AP^m_r	AP^m_c	AP^m_f	$m\mathbf{AP}$
(1)	Baseline	19.7	1.0	18.0	29.9	22.1
(2)	RFS† [15]	23.0	10.4	24.4	29.5	26.5
(3)	RFS-cls† [15]	23.0	9.5	24.1	29.4	25.7
(4)	Focal loss‡ [26]	10.4	0.8	5.1	20.6	11.2
(5)	Focal loss-cls‡ [26]	17.2	0.9	14.2	27.8	18.7
(6)	EQ loss† [27]	21.6	4.1	22.1	28.4	24.6
(7)	seesaw loss‡ [28]	21.4	2.5	21.7	29.6	24.3
(8)	seesaw loss-cos‡ [28]	24.8	13.6	25.9	30.5	27.7
(9)	CChead† [29]	21.5	13.2	18.3	29.6	24.5
(10)	BAGS† [25]	25.6	17.2	25.7	29.3	28.8
(11)	De-confound-TDE† [20]	25.0	14.2	24.6	30.0	28.2
(12)	MEC	**27.6**	**18.9**	**28.2**	**31.1**	**30.5**

† Reproducing the results from their official code.
‡ Re-implementing by us.

Table 2. Compare the performance of the proposed MEC framework with other long-tailed approaches COCO-LT.

Model	$m\mathbf{AP}^m$	AP_{50}	AP_{75}	AP^m_r	AP^m_c	AP^m_f	$m\mathbf{AP}$
Baseline	14.0	24.5	14.3	3.8	17.7	20.0	16.7
EQLoss† [27]	12.8	22.1	13.0	3.3	14.2	**20.3**	15.2
SeesawLoss‡ [28]	14.1	24.6	14.3	3.8	**18.2**	19.8	16.8
MEC	**14.5**	**24.9**	**14.5**	**5.0**	18.0	19.8	**17.1**

3.1. Evaluation on LVIS

As depicted in Table 1, MEC outperforms not only the conventional re-sampling/re-weighting approaches (models (2-5)) but also the recently proposed gradient re-weighting methods, such as EQ Loss (model (6)) and seesaw loss (models (7-8)), by a significant margin, especially on rare and common categories. It is noteworthy that our proposed method, MEC, surpasses the current state-of-the-art methods (models (10, 11)) by 2.0 and 2.6 AP, respectively. This comparison results not only demonstrate that MEC can obtain a better representation but also suggest that addressing imbalanced detection problems is necessary.

3.2. Evaluation on COCO-LT

To further assess the generalization of our method, we created a COCO-LT dataset with a similar long-tailed distribution as LVIS, by sampling images and annotations from the COCO [13] dataset. The statistical results can be found in our supplementary material. As shown in Table 2, we observed that the proposed MEC outperforms the baseline, especially on rare categories. These results confirm the effectiveness of MEC for long-tailed object detection tasks.

3.3. Ablation Studies

In order to analyze the proposed multi-expert loss and dynamic ensemble mechanism, we conducted two ablation studies in this section. The first study evaluates the adaptability of the multi-expert loss to different backbones and frameworks. Table 3 shows that the multi-expert loss can effectively adapt to different multi-stage architectures such as Cascade R-CNN and Hybrid Task Cascade (HTC). We also incorporated the multi-expert loss with recently proposed long-tailed classification methods and found that it consistently improves performance, especially on rare and common categories. These results demonstrate that the multi-expert loss does not conflict with other state-of-the-art methods. The second ablation study focused on parsing the proposed approaches. Table 4 compares the proposed approaches and shows that the baseline with SE echoes the observation in Figure 1(a), while the dynamic ensemble mechanism can reach the potential of multi-expert loss training. These experiments provide evidence that the MEC is effective and robust.

Table 3. The performance about different backbones with multi-expert loss and other long-tailed methods on LVIS v1.0.

ID	Model	Backbone	Multi-expert Loss	$m\mathbf{AP}^m$	AP^m_r	AP^m_c	AP^m_f	$m\mathbf{AP}$
(1)	Cascade R-CNN† [11]	R50-FPN	×	19.7	1.0	18.0	**29.9**	22.1
(2)			✓	**23.8**	**6.9**	**23.7**	31.3	**26.5**
(3)	Cascade R-CNN† [11]	R101-FPN	×	21.6	1.8	20.8	31.1	24.2
(4)			✓	**25.0**	**7.6**	**25.1**	**32.6**	**27.9**
(5)	HTC† [12]	R50-FPN	×	20.4	1.4	19.3	31.3	23.3
(6)			✓	**24.3**	**7.2**	**24.3**	**32.9**	**27.1**
(7)	HTC† [12]	R101-FPN	×	22.1	1.5	20.7	32.7	24.2
(8)			✓	**26.4**	**8.1**	**26.6**	**34.2**	**28.6**
(9)	SeesawLoss‡ [28]	R50-FPN	×	21.4	2.5	21.7	29.6	24.3
(10)			✓	**23.8**	**6.6**	**24.2**	**31.7**	**26.9**
(11)	Cos-Norm‡ [30]	R50-FPN	×	22.6	6.5	22.0	**30.3**	25.6
(12)			✓	**27.4**	**18.7**	**27.4**	31.1	**30.4**
(13)	De-confound† [20]	R50-FPN	×	25.0	14.2	24.6	30.1	28.2
(14)			✓	**26.8**	**17.3**	**26.3**	**31.5**	**30.1**

† Reproducing the results from their official code.
‡ Re-implementing by us.

Table 4. The ablation study of the proposed multi-expert loss (ME), dynamic ensemble mechanism (DE), and sparse ensemble mechanism (SE).

Model	ME	DE	SE	$m\mathbf{AP}^m$	AP^m_r	AP^m_c	AP^m_f	$m\mathbf{AP}$
R50-FPN				19.7	1.0	18.0	29.9	22.1
			✓	22.5	5.2	21.1	31.6	25.0
	✓			23.8	6.9	23.7	31.3	26.5
	✓		✓	25.2	8.7	25.4	**32.2**	27.7
	✓	✓		**27.6**	**18.9**	**28.2**	31.1	**30.5**

4. CONCLUSIONS

We propose a novel Multi-Expert Cascade (MEC) approach to enhancing the multi-stage detector framework for long-tailed object detection. By tackling the imbalanced detection issue and investigating the intricate relationship between the IoU thresholds and data frequency, MEC effectively improves detection performance, especially for rare and common categories. The experimental results demonstrate the effectiveness and robustness of the proposed MEC. The MEC pioneers the concept that modeling the relationship between categories and their corresponding frequencies is essential for advancing long-tailed object detection.

Acknowledgements. This work was supported in part by the NSTC grants 111-2221-E-001-015-MY3, 111-2221-E-001-011-MY2 and 111-2634-F-007-010 of Taiwan. We are grateful to National Center for High-performance Computing for providing computational resources and facilities.

5. REFERENCES

[1] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 770–778.

[3] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg, "SSD: single shot multibox detector," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, 2016, pp. 21–37.

[4] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 779–788.

[5] Joseph Redmon and Ali Farhadi, "YOLO9000: better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 6517–6525.

[6] Joseph Redmon and Ali Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018.

[7] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, 2014, pp. 580–587.

[8] Ross B. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015, pp. 1440–1448.

[9] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 2015, pp. 91–99.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick, "Mask R-CNN," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017, pp. 2980–2988.

[11] Zhaowei Cai and Nuno Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[12] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al., "Hybrid task cascade for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 4974–4983.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014.

[14] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[15] Agrim Gupta, Piotr Dollar, and Ross Girshick, "LVIS: A dataset for large vocabulary instance segmentation," in *CVPR*, 2019.

[16] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen, "BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9719–9728.

[17] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[18] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong, "Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020, pp. 7607–7616.

[19] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10991–11000.

[20] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang, "Long-tailed classification by keeping the good and removing the bad momentum causal effect," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[21] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph, "Simple copy-paste is a strong data augmentation method for instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 2918–2928.

[22] Bo Li, Yongqiang Yao, Jingru Tan, Gang Zhang, Fengwei Yu, Jianwei Lu, and Ye Luo, "Equalized focal loss for dense long-tailed object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 6990–6999.

[23] Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, and Shu Kong, "Long-tailed recognition via weight balancing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 6897–6907.

[24] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[25] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020, pp. 10988–10997.

[26] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017, pp. 2999–3007.

[27] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan, "Equalization loss for long-tailed object recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020, pp. 11659–11668.

[28] 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," *arXiv preprint arXiv:2008.10032*, 2020.

[29] Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Jun Hao Liew, Sheng Tang, Steven Hoi, and Jiashi Feng, "Classification calibration for long-tail instance segmentation," *arXiv preprint arXiv:1910.13081*, 2019.

[30] Spyros Gidaris and Nikos Komodakis, "Dynamic few-shot visual learning without forgetting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4367–4375.