# Learning with Free Object Segments for Long-Tailed Instance Segmentation

Cheng Zhang Tai-Yu Pan Tianle Chen Jike Zhong Wenjin Fu Wei-Lun Chao

The Ohio State University, Columbus, OH, USA

## Abstract

*In this paper, we explore the possibility to increase the training examples without laborious data collection and annotation for long-tailed instance segmentation. We find that an abundance of instance segments can potentially be obtained freely from object-centric images, according to two insights: (i) an object-centric image usually contains one salient object in a simple background; (ii) objects from the same class often share similar appearances or similar contrasts to the background. Motivated by these insights, we propose a simple and scalable framework* FREESEG *for extracting and leveraging these "free" object segments to facilitate model training. Concretely, we investigate the similarity among object-centric images of the same class to propose candidate segments of foreground instances, followed by a novel ranking of segment quality. The resulting high-quality object segments can then be used to augment the existing long-tailed datasets,* e.g.*, by copying and pasting the segments onto the original training images. Extensive experiments show that* FREESEG *yields substantial improvements on top of strong baselines and achieves state-of-the-art accuracy for segmenting rare object categories.*

## 1. Introduction

Recent years have witnessed an unprecedented breakthrough in common object detection and instance segmentation [3,5,7,12,16,30]. Yet, when it comes to rare, less commonly seen objects, there is a drastic performance drop due to insufficient training examples [6, 18, 28]. This challenge has attracted significant attention lately in how to learn a detection or segmentation model given labeled data of a "long-tailed" distribution across classes [6, 8, 11, 14, 19–24, 29].

In this paper, *we investigate the possibility of obtaining more labeled instances (*i.e.*, instance segments of objects) under a minimal cost, especially for rare objects.* We build upon the recent observation in [26] — many objects do not appear frequently enough in complex scenes, but are found frequently alone in object-centric images — to acquire an abundance of object-centric images (*e.g.*, ImageNet [2] or Google images) for rare classes. In general, object-centric images mostly contain one salient object in a relatively simple background than scene-centric images like those in MSCOCO [12]. Moreover, objects of the same class usually share similar appearances, shapes, or contrasts to the background (See Figure 1 for an example). These properties open up the opportunity to discover object segments almost *freely* from object-centric images of the same class — by exploring their common salient regions.

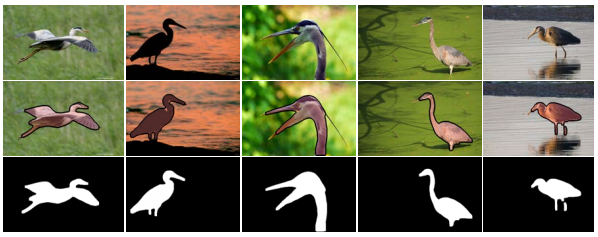

Figure 1. **Motivation of our approach FREESEG.** We sample a rare classes, *heron* from LVIS v1 [6], and retrieve object-centric images (the upper row) from the ImageNet dataset [17]. We then show the discovered object segments (the middle row) and binary masks (the bottom row) by FREESEG. The abundant object segments have diverse appearances and poses and can be effectively used to improve the instance segmentation.

To this end, we propose a framework named FREESEG (Free Object Segments) to take advantage of these properties. We first extract the common foreground regions from object-centric images of the same class by off-the-shelf co-segmentation models [27]. However, directly using all of these regions, mixed with false positive and noisy segments, would inevitably introduce a great amount of noise to the downstream tasks. To address this, we propose a novel *segment ranking* approach to mine the most reliable and high-confident object segments.

One naive way to leverage these instance segments is to directly train on the object-centric images. Nevertheless, these objects mostly show up alone in simple backgrounds makes them somewhat too simple for the model. We therefore place these object segments in the context of complex scene-centric images, via simple copy-paste augmentation [4]. Unlike [4], which merely pastes ground-truth

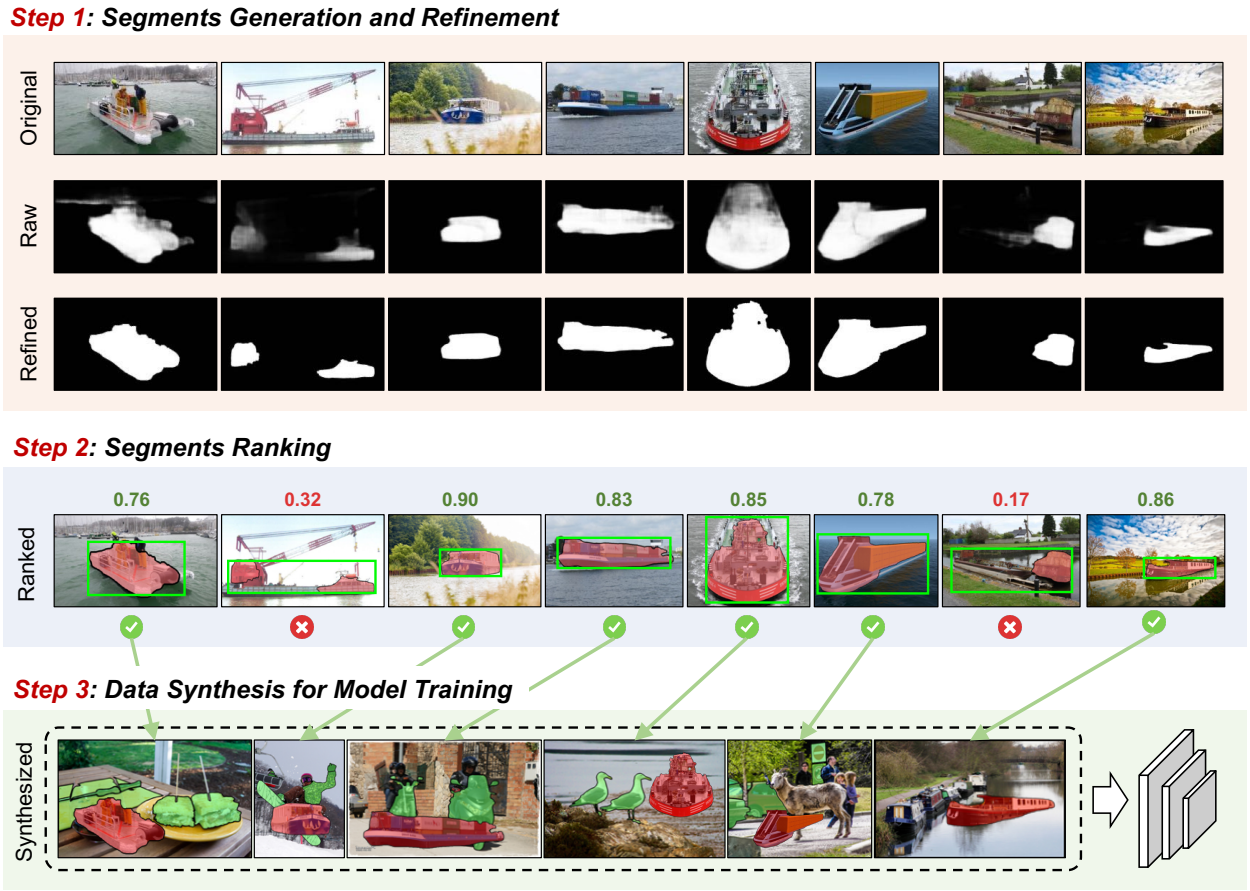


Figure 2. **Illustration of the FREESEG pipeline.** We show a rare class *Barge* in LVIS v1 [6] as the example. We first perform image co-segmentation on top of the object-centric images of *Barge* (outside LVIS v1) to obtain raw object segments, followed by segments refinement. The segments are then scored by a learned ranker (the green boxes in step 2) such that only the high-quality ones would be used for augmenting data for model training. Finally, we randomly paste the selected object segments (red) onto the original scene-centric images of LVIS v1 to improve the long-tailed instance segmentation. Green segments indicate the original objects in scene-centric images.

segments from one image to another to increase the *context diversity*, our approach brings the best of abundant free object segments to increase the *appearance diversity*.

In summary, our **main contributions** are:

- We demonstrate the possibility to increase the number of training examples for instance segmentation without laborious pixel-level data collection and annotation.
- We propose a simple and scalable pipeline for discovering, extracting, and leveraging free object foreground segments to facilitate long-tailed instance segmentation.
- Our FREESEG framework achieves state-of-the-art performance on the challenging LVIS dataset and demonstrates a strong compatibility with existing works.

## 2. FREESEG for Data Augmentation

Figure 2 illustrates the pipeline, which consists of three major steps: (i) segment generation and refinement, (ii) segment ranking, and (iii) data synthesis for model training.

### 2.1. Generating Object Segments

**Bootstrap object-centric images collection.** We first collect object-centric images for each class of interest. As discussed in [26], we use the unique WordNet synset ID [13] to match the categories between ImageNet-22K [17] and LVIS v1 [6]. We are able to match 997 LVIS classes ($1,242,180$ images from ImageNet). Because ImageNet images are nearly balanced by design (with around 1K images/class), the imbalance situation in LVIS can be largely reduced. We further retrieve images via Google by querying with class names provided by LVIS. Such a search returns hundreds of iconic images and we take top 500 for each class.

**Segments generation and refinement.** Given images from the same class, we then apply image co-segmentation techniques [27] to extract their common foreground regions. We also threshold [9,10] the map and apply erosion and dilation to smooth the boundary. Finally, we remove small, likely false positive segments by only keeping the largest connected component in the binary map (Step 1 of Figure 2).

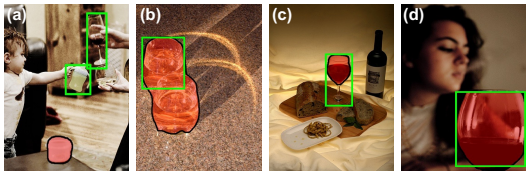


Figure 3. **Comparison of metrics for ranking segments.** We show four examples of the class *wine glasses*. The red masks are by our method; green boxes are by LORE. In (a) and (d), IoU ranks the segments well, when the box locations are precise. However, in (b), the poor box location leads to a small IoU, even if the segment is precise. In (c), IoU fails due to the specific shape of *wine glasses*, even if the segment is precise. FREESEG score is able to take all above into account to faithfully rank segments.

### 2.2. Learning to Rank the Segments

**Ranking by learning a classifier.** How can we determine if the segment truly covers the target object? Here, we take one intuition [26]: *if a segment covers the target object, then by removing it from the image, an image classifier*[1] *will unlikely classify the manipulated image correctly.* The authors in [26] developed "localization by region removal (LORE)", which sequentially removes bounding box regions from an image till the image classifier fails to predict the right class. Those removed bounding boxes are then treated as pseudo bounding boxes for the target object class. We thus adopt the idea of LORE to rank our object segments. Instead of removing the discovered segments and checking the classifier's failure, we directly compare our object segments to the bounding boxes selected by LORE. *In essence, if the LORE boxes and our segments are highly overlapped, then the segments are considered high-quality.*

**Segments ranking.** The most common way to characterize the overlap/agreement between two masks/boxes is intersection over union (IoU). However, this metric is not suitable as: (i) both boxes and segments may be noisy, and simply measuring the IoU between them fails to rank good segments when the boxes are poor; (ii) object shapes are not always convex, and IoU may underestimate the agreement. As shown in Figure 3, IoU fails to recall true positives.

We make one mild assumption: either the object box or the segment is trustable, and introduce two metrics: intersection over bounding box (IoB) and intersection over mask (IoM). While they share the same numerator with intersection over union (IoU), they have different denominators. IoM implies that the bounding box is precise and measures how much portion of the mask is inside the box, and vice versa for IoB. We take both into account by averaging them as FREESEG score. We jointly use (i) FREESEG score and (ii) the classifier's relative confidence drop for the target class before and after LORE box removal, to rank the object-centric images and their co-segmentation segments. We keep those with both scores larger than 0.5 as the high-quality segments. As shown in Figure 2 (Step 2), our method effectively keeps the good segments in the pool.

[1]We have image labels for object-centric images and thus we can train an image classifier upon them.

### 2.3. Putting the Segments in the Context

We adopt copy-paste [4] to randomly (i) sample several object-centric images, (ii) re-scale and horizontally flip the object segments, and (iii) paste them onto the scene-centric images from the original training set. The synthesized images can be used to improve model training (Step 3 in Figure 2). See the supplementary material for more details.

## 3. Experiments

### 3.1. Setup

**Dataset and evaluation metrics.** We validate FREESEG on LVIS v1 instance segmentation benchmark [6]. The categories follow a long-tailed distribution and are divided into three groups: rare (1-10 images), common (11-100 images), and frequent (>100 images). We adopt the standard mean average precision (AP) metric. We denote the AP for rare, common, and frequent classes as $AP_r$, $AP_c$, and $AP_f$, respectively. We also report AP for bounding boxes (*i.e.*, $AP^b$), predicted by the same instance segmentation models.

**Base models.** We using two base models for instance segmentation, *i.e.*, Mask R-CNN [7] and MosaicOS [26]. Mask R-CNN is trained with the LVIS v1 training set, following the standard training procedure [6]. MosaicOS [26] is one of the state-of-the-art models which is further pre-trained with balanced object-centric images from ImageNet-22K and Google Images.

**Training and optimization.** Given the base instance segmentation model, we first fine-tune the model for 90K iterations with FREESEG segments, using all the loss terms in Mask R-CNN. We then fine-tune the model again for another 90K iterations using the original LVIS training images. The rationale of training with multiple stages is to prevent the augmented instances from dominating the training process and it is shown to be effective in [26].

### 3.2. Results and Analyses

**Main results.** We compare to state-of-the-art methods for long-tailed instance segmentation in Table 1. The proposed FREESEG method achieves comparable or even better results, especially for rare categories. We further apply post-processing calibration [15] on top the model trained with FREESEG and show results in Table 1 (FREESEG $\star$). Surprisingly, FREESEG can boost the performance of rare

Table 1. **State-of-the-art comparison on LVIS v1 instance segmentation.** FREESEG are initialized with MosaicOS [26] as the base model. 2×: Seesaw applies a stronger 2× training schedule while other methods are with 1× schedule. ⋆: with post-processing calibration introduced by [15].

| Method | AP | $AP_r$ | $AP_c$ | $AP_f$ | $AP^b$ |
|---|---|---|---|---|---|
| *Mask R-CNN with ResNet-50 FPN* | | | | | |
| RFS [6] | 22.58 | 12.30 | 21.28 | 28.55 | 23.25 |
| BaGS [11] | 23.10 | 13.10 | 22.50 | 28.20 | 25.76 |
| RIO [1] | 23.70 | 15.20 | 22.50 | 28.80 | 24.10 |
| EQL v2 [19] | 23.70 | 14.90 | 22.80 | 28.60 | 24.20 |
| FASA [25] | 24.10 | 17.30 | 22.90 | 28.50 | – |
| Seesaw [21]$^{2\times}$ | 26.40 | 19.60 | 26.10 | 29.80 | 27.40 |
| MosaicOS [26] | 24.45 | 18.17 | 23.00 | 28.83 | 25.05 |
| w/ FREESEG | 25.19 | ***20.23*** | 23.80 | 28.92 | 25.98 |
| MosaicOS [26] ⋆ | 26.76 | 23.86 | 25.82 | 29.10 | 27.77 |
| w/ FREESEG ⋆ | 27.34 | ***25.11*** | 26.29 | 29.49 | 28.47 |
| *Mask R-CNN with ResNet-101 FPN* | | | | | |
| RFS [6] | 24.82 | 15.18 | 23.71 | 30.31 | 25.45 |
| FASA [25] | 26.30 | 19.10 | 25.40 | 30.60 | – |
| Seesaw [21]$^{2\times}$ | 28.10 | 20.00 | 28.00 | 31.90 | 28.90 |
| MosaicOS [26] | 26.73 | 20.52 | 25.78 | 30.53 | 24.41 |
| w/ FREESEG | 27.54 | ***23.00*** | 26.48 | 30.72 | 28.63 |
| MosaicOS [26] ⋆ | 29.03 | 26.38 | 28.15 | 31.19 | 29.96 |
| w/ FREESEG ⋆ | 29.72 | ***28.69*** | 28.67 | 31.34 | 31.11 |
| *Mask R-CNN with ResNeXt-101 FPN* | | | | | |
| RFS [6] | 26.67 | 17.60 | 25.58 | 31.89 | 27.35 |
| MosaicOS [26] | 28.29 | 21.75 | 27.22 | 32.35 | 28.85 |
| w/ FREESEG | 28.86 | ***23.34*** | 27.77 | 32.49 | 29.98 |
| MosaicOS [26] ⋆ | 29.81 | 25.73 | 28.92 | 32.59 | 30.56 |
| w/ FREESEG ⋆ | 30.37 | ***26.43*** | 29.63 | 32.92 | 31.81 |

classes to be similar to common classes. This indicates that by introducing more while not so perfect training instances, FREESEG dramatically overcomes the long-tailed problem.

**Does segments ranking help?** We conduct experiments with and without ranking on object segments in Table 2. We are able to collect $1,830$K segments from ImageNet-22k and Google Images, while only 966K of them are left after filtering with FREESEG. While both versions outperform the baseline models, segments ranking does help more (row 4 *vs.* row 2 in Table 2). Since filtering by ranking gives higher quality but fewer data than that without ranking, we surmise that this somehow limits the gain. Thus, we randomly sample segments to have the same number as those after filtering by ranking, and train a model with them. We see a bigger gain by filtering (row 4 *vs.* row 3 in Table 2), justifying the effectiveness of ranking.

**Ranking metrics.** We compare the results using different ranking metrics in Table 3. FREESEG score can take different scenarios into account and successfully select confident segments from noisy ones.

**Comparison to pasting ground-truth segments.** We compare pasting ground-truth [4] and FREESEG object segments in Table 4. FREESEG achieves consistently gains against the baseline models and outperforms vanilla copy-paste [4]. Note that FREESEG is more effective when the baseline is already re-balanced (*e.g.*, MosaicOS in Table 4 bottom), while GT-only can hardly improve upon it due to the lack of training examples. Furthermore, by learning with copy-paste from both sources, the gain can be even larger. These demonstrate that the appearance diversity of objects is also the key to improve instance segmentation.

Table 2. **Ablation study on segments ranking.** We evaluate the model trained with and without ranking (Rank) mechanism or randomly (Rand) sampled the segments by FREESEG.

| Method | Rand | Rank | #Img | AP | $AP_r$ | $AP_c$ | $AP_f$ |
|---|---|---|---|---|---|---|---|
| MosaicOS [26] | | | | 24.45 | 18.17 | 23.00 | 28.83 |
| | | | 1,830K | 24.87 | 19.13 | 23.55 | 28.86 |
| w/ FREESEG | ✓ | | 966K | 24.50 | 18.68 | 23.18 | 28.52 |
| | | ✓ | 966K | **25.19** | **20.23** | **23.80** | **28.92** |

Table 3. **Analysis on different object segments ranking metrics.**

| Method | Metrics | AP | $AP_r$ | $AP_c$ | $AP_f$ |
|---|---|---|---|---|---|
| MosaicOS [26] | – | 24.45 | 18.17 | 23.00 | 28.83 |
| w/ FREESEG | IoU | 24.74 | 19.04 | 23.58 | 28.53 |
| | IoB | 24.69 | 18.41 | 23.58 | 28.70 |
| | IoM | 24.56 | 18.62 | 23.14 | 28.74 |
| | FREESEG | **25.19** | **20.23** | **23.80** | **28.92** |

Table 4. **Comparison of pasting ground-truth (GT) object segments [4] and FREESEG.**

| Method | GT [4] | FREESEG | AP | $AP_r$ | $AP_c$ | $AP_f$ |
|---|---|---|---|---|---|---|
| Mask R-CNN [6]† | | | 22.58 | 12.30 | 21.28 | 28.55 |
| | ✓ | | 24.06 | 17.00 | 22.62 | 28.77 |
| | | ✓ | 24.28 | 17.68 | 22.79 | 28.83 |
| | ✓ | ✓ | 24.74 | ***18.80*** | 23.38 | 28.86 |
| MosaicOS [26]† | | | 24.45 | 18.17 | 23.00 | 28.83 |
| | ✓ | | 24.57 | 18.63 | 23.31 | 28.59 |
| | | ✓ | 25.19 | 20.23 | 23.80 | 28.92 |
| | ✓ | ✓ | 25.36 | ***20.72*** | 24.00 | 28.92 |

**Additional results.** Please see supplementary material, including the analysis on multi-stage training, effects of data source, additional evaluation metric, qualitative results, etc.

## 4. Conclusion

We propose a scalable framework FREESEG to take the best usage of object-centric images to facilitate long-tailed instance segmentation. We show that, with the underlying properties of object-centric images, simple co-segmentation with proper ranking can result in high-quality instance segments to largely increase the labeled training instances. We expect our approach to serve as a strong baseline for this task: for future work to build upon and take advantage of.

**Acknowledgments** This research is partially supported by NSF IIS-2107077, NSF OAC-2118240, NSF OAC-2112606, the OSU GI Development fund, and the OSU CCTS Pilot grant. We are thankful for the generous support of the computational resources by the Ohio Supercomputer Center and AWS Cloud Credits for Research.

## References

[1] Nadine Chang, Zhiding Yu, Yu-Xiong Wang, Anima Anandkumar, Sanja Fidler, and Jose M Alvarez. Image-level or object-level? a tale of two resampling strategies for long-tailed detection. In *ICML*, 2021. 4

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[3] 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 (IJCV)*, 88(2):303–338, 2010. 1

[4] 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 *CVPR*, 2021. 1, 3, 4

[5] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1

[6] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 1, 2, 3, 4

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 1, 3

[8] Xinting Hu, Yi Jiang, Kaihua Tang, Jingyuan Chen, Chunyan Miao, and Hanwang Zhang. Learning to segment the tail. In *CVPR*, 2020. 1

[9] CH Li and Peter Kwong-Shun Tam. An iterative algorithm for minimum cross entropy thresholding. *Pattern Recognition Letters (PRL)*, 19(8):771–776, 1998. 2

[10] Chun Hung Li and CK Lee. Minimum cross entropy thresholding. *Pattern recognition*, 26(4):617–625, 1993. 2

[11] 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 *CVPR*, 2020. 1, 4

[12] 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. 1

[13] George A Miller. WordNet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 2

[14] Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. Imbalance problems in object detection: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 1

[15] Tai-Yu Pan, Cheng Zhang, Yandong Li, Hexiang Hu, Dong Xuan, Soravit Changpinyo, Boqing Gong, and Wei-Lun Chao. On model calibration for long-tailed object detection and instance segmentation. In *NeurIPS*, 2021. 3, 4

[16] 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 (TPAMI)*, 39(6):1137–1149, 2016. 1

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 1, 2

[18] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019. 1

[19] 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 *CVPR*, 2021. 1, 4

[20] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *CVPR*, 2020. 1

[21] 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 *CVPR*, 2021. 1, 4

[22] Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Junhao Liew, Sheng Tang, Steven Hoi, and Jiashi Feng. The devil is in classification: A simple framework for long-tail instance segmentation. In *ECCV*, 2020. 1

[23] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. In *ICML*, 2020. 1

[24] Jialian Wu, Liangchen Song, Tiancai Wang, Qian Zhang, and Junsong Yuan. Forest R-CNN: Large-vocabulary long-tailed object detection and instance segmentation. In *ACM MM*, 2020. 1

[25] Yuhang Zang, Chen Huang, and Chen Change Loy. FASA: Feature augmentation and sampling adaptation for long-tailed instance segmentation. In *ICCV*, 2021. 4

[26] Cheng Zhang, Tai-Yu Pan, Yandong Li, Hexiang Hu, Dong Xuan, Soravit Changpinyo, Boqing Gong, and Wei-Lun Chao. MosaicOS: a simple and effective use of object-centric images for long-tailed object detection. In *ICCV*, 2021. 1, 2, 3, 4

[27] Kaihua Zhang, Jin Chen, Bo Liu, and Qingshan Liu. Deep object co-segmentation via spatial-semantic network modulation. In *AAAI*, 2020. 1, 2

[28] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*, 2021. 1

[29] Xiangxin Zhu, Dragomir Anguelov, and Deva Ramanan. Capturing long-tail distributions of object subcategories. In *CVPR*, 2014. 1

[30] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*, 2019. 1