This ACCV 2022 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ACCV 2022
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/accv

Boosting Dense Long-Tailed Object Detection from Data-Centric View

Weichen Xu, Jian Cao(✉), Tianhao Fu, Hongyi Yao, and Yuan Wang

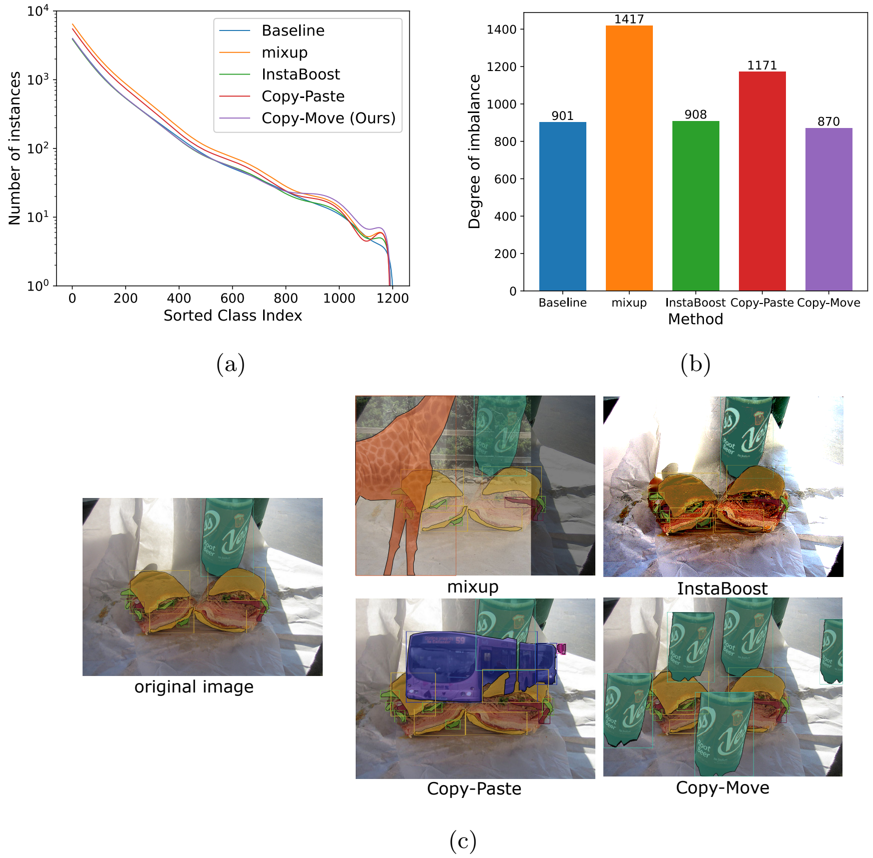
Peking University, Beijing, China
{xuweichen1999,tianhaofu1,yhy}@stu.pku.edu.cn, {caojian}@ss.pku.edu.cn

Abstract. Several re-sampling and re-weighting approaches have been proposed in recent literature to address long-tailed object detection. However, state-of-the-art approaches still struggle on the rare class. From data-centric view, this is due to few training data of the rare class and data imbalance. Some data augmentations which could generate more training data perform well in general object detection, while they are hardly leveraged in long-tailed object detection. We reveal that the real culprit lies in the fact that data imbalance has not been alleviated or even intensified. In this paper, we propose REDet: a rare data centric detection framework which could simultaneously generate training data of the rare class and deal with data imbalance. Our REDet contains data operations at two levels. At the instance-level, Copy-Move data augmentation could independently rebalance the number of instances of different classes according to their rarity. Specifically, we copy instances of the rare class in an image and then move them to other locations in the same image. At the anchor-level, to generate more supervision for the rare class within a reasonable range, we propose Long-Tailed Training Sample Selection (LTTSS) to dynamically determine the corresponding positive samples for each instance based on the rarity of the class. Comprehensive experiments performed on the challenging LVIS v1 dataset demonstrate the effectiveness of our proposed approach. We achieve an overall 30.2% AP and obtain significant performance improvements on the rare class.

1 Introduction

In real-world scenarios, training data generally exhibit a long-tailed class distribution, where a small number of classes have a large amount of data, but others have only a small amount of data [1]. Long-tailed object detection is receiving increasing attention because of the need for realistic scenarios.

Some existing approaches deal with this task by data re-sampling [2,3,4,5] or loss re-weighting [6,7,8,9]. Specifically, re-sampling approaches increase rare class instances by performing image-level resampling in the dataset, which is effective when a certain amount of image-level training data contains rare class instances. The re-weighting approaches increase the contribution of the rare class to the gradient by modifying the loss function, which in turn increases the focus on



(a) (b)

(c)

Fig. 1: (a) The number of instances of baseline and various data augmentations on LVIS v1 [2] train split. Classes' indices are sorted by instance counts of baseline. (b) The degree of imbalance between baseline and various data augmentations on LVIS v1 [2] train split. (c) Visualization of the original image and various data augmentations. For display, the image augmented by Copy-Paste is scaled to the same size as the other images.

the rare class. In addition, many remarkable efforts have focused on incremental learning [10], data augmentation [11], and decoupled learning [4,12,13].

However, state-of-the-art approaches still struggle on the rare class. In fact, the poor performance of current SOTA long-tailed detection methods is caused by the dataset quality itself. An intuitive example is that if we convert a dataset from the long-tailed dataset into a balanced dataset, there will be no such problem as long-tailed object detection. Thus, the benefits of improving the model or the loss function are far less obvious than improving the dataset directly. Recently, Data-Centric AI [14] has been a scorching research topic. The main idea is usually to do a series of operations on the data so that the gradient is updated in a more optimal direction when updating the model parameters. If we rethink long-tailed object detection from data-centric view, we could find that there are two main difficulties [1]: (1). lack of the rare class instances leads to poor performance; (2). drastic data imbalance makes the performance of the rare class affected by the frequent class. Fig. 1(a) shows the number of instances of each class in LVIS [2] dataset. It can be seen that the task is challenging due to the above two main difficulties. Some data augmentations try to deal with lacking of the rare class, while these methods still cannot resolve the data imbalance.

Fig. 1(a) shows the impact of several common data augmentations in object detection on the number of instances. It can be seen that mixup [15], InstaBoost [16], and Copy-Paste [11] are relatively crude data augmentations for long-tailed object detection. They use the same rules for all classes and do not consider the rarity of the class. They cannot solve the label co-occurrence problem. The frequent class is augmented at the same time. Here, We use the standard deviation of the number of instances in all classes to represent the degree of imbalance. Fig. 1(b) illustrates the degree of imbalance with and without various data augmentations. It can be seen that degree of imbalance has not been alleviated or even intensified due to blindly increasing the number of instances in all classes.

Therefore, we can generate more rare class data by modifying the data in a more refined way. So we propose REDet: a rare data centric detection framework that could bridge the gap of handling long-tailed distribution data at the instance and anchor levels. Specifically, at the instance-level, we propose Copy-Move data augmentation, which introduces information about the long-tailed distribution into the data augmentation. Instances are copied and moved to other locations in the same image according to the rarity of each class. Without destroying the semantic information of the image, we increase the diversity of the rare class in the dataset to alleviate the lack of rare class instances and data imbalance. At the anchor-level, we propose Long-Tailed Training Sample Selection (LTTSS) to dynamically determine the corresponding positive samples for each instance based on the rarity of its class. Our approach has the lowest degree of imbalance compared to other data augmentations. It yields a new state-of-the-art and can be well applied to existing dense long-tailed object detection pipelines.

To sum up, our key contributions can be summarized as follows:

- We think about long-tailed object detection from data-centric view and propose degree of imbalance to evaluate several existing data augmentations.
- We propose REDet: a rare data centric detection framework in which Copy-Move and LTTSS work collaboratively to promote the instances balance and positive samples balance while increasing training data.
- Extensive experiments on the challenging LVIS dataset demonstrate the effectiveness of the proposed approach. Our approach achieves state-of-the-art results on LVIS by introducing long-tailed information in data augmentation and training sample selection.

2 Related Work

General Object Detection. Object detection approaches have achieved immense success in recent years, benefiting from the powerful classification ability of convolutional neural networks (CNN) [17,18,19]. Advanced object detectors can be categorized into two-stage and one-stage approaches. Two-stage approaches [20,21,22] first generate coarse proposals through region proposal network (RPN). Then, these proposals are further refined for accurate classification and bounding box regression. One-stage approaches [23,24,25,26] make predictions directly on the dense anchors or points without generating coarse proposals. In practice, one-stage detectors are more widely used in real-world scenarios. But the performance of the general object detectors degrades dramatically when it comes to the long-tailed distribution of data [6].

Long-Tailed Object Detection. Long-tailed object detection is more complex than general object detection due to the extreme data imbalance. It is receiving increasing attention [1]. One classic solution to this problem is loss re-weighting. The basic idea of the re-weighting method is to assign different weights to the training samples based on the rarity of the class. Tan et al. [6] proposed the equalization loss (EQL) that ignored the negative gradients from frequent samples. Seesaw loss [7] proposed compensation factor to avoid false positives of the rare class. EQLv2 [8] rethought the essential role of samples in the classification branch and adopted a gradient-guided mechanism to reweight the loss of each class. Li et al. [9] proposed the equalized focal loss (EFL) that rebalanced the loss contribution of positive and negative samples in one-stage detectors. Another useful solution is the re-sampling strategy. Repeat factor sampling (RFS) [2] over-sampled images containing rare classes to balance the data distribution at the image-level. At the instance-level, Forest R-CNN [3] set a higher non-maximum suppression (NMS) threshold for the rare class to get more proposals. Other works [4,5] used bi-level class balanced sampler or memory-augmented sampler to implement data resampling. However, both re-weighting and re-sampling approaches still struggle on the rare class due to the lack of consideration of the long-tailed distribution in dataset.

Data Augmentations. Data augmentations such as CutMix [27], InstaBoost [16], and Mosaic [28] can significantly boost object detection performance. However, as a simple technique, data augmentation is rarely discussed in long-tailed object detection. MiSLAS [29] proposed to use data mixup to enhance representation learning in long-tailed image classification. Ghiasi et al. [11] demonstrated that the simple mechanism of pasting objects randomly was good enough for the long-tailed instance segmentation task. Zhang et al. [30] addressed the data scarcity issue by augmenting the feature space, especially for the rare class. Simply using existing augmentation techniques for improving long-tailed object detection performance is unfavorable, which will lead to the problem of label co-occurrence. Specifically, frequent class labels frequently appear with rare class labels during data augmentation. Thus, the frequent class would be augmented more, which may bias the degree of imbalance. Instead, we dynamically increase the number of instances and positive samples according to the rarity of the class, which can solve the above problem well and have a higher data validity.

3 Method

The rare data centric detection framework we proposed is based on one-stage object detection network such as RetinaNet [23]. Fig. 2 is an overview of the proposed REDet, which shows that our Copy-Move data augmentation is inserted before network, and LTTSS is used to select positive samples after network. For instances of the rare class, the copy times and the long-tailed scaling factor are calculated for Copy-Move and LTTSS, respectively.

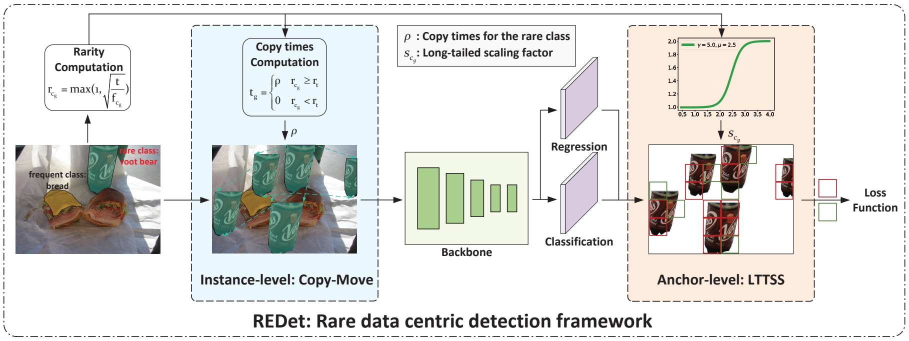

Rarity Computation
$r_{c_g} = \max(1, \sqrt{\frac{t}{f_{c_g}}})$
frequent class: bread
rare class: root bear
Copy times Computation
$t_g = \begin{cases} \rho & r_{c_g} \geq r_t \\ 0 & r_{c_g} < r_t \end{cases}$
ρ
Instance-level: Copy-Move
ρ : Copy times for the rare class
s_{c_g} : Long-tailed scaling factor
Backbone
Regression
Classification
$\gamma = 5.0, \mu = 2.5$
2.0
1.8
1.6
1.4
1.2
1.0
0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
s_{c_g}
Anchor-level: LTTSS
Loss Function
REDet: Rare data centric detection framework

Fig. 2: Architecture overview of REDet. Our proposed method contains two main components: Copy-Move data augmentation at the instance-level and LTTSS at the anchor-level. In LTTSS, the red anchors represent original positive samples, and the green anchors represent the newly obtained positive samples after introducing the long-tailed information.

3.1 Instance-Level: Copy-Move

The existing instance-level data augmentation approach [11] selected a random subset of instances from one image and pasted them onto another image. However, images generated in this way could look very different from authentic images in terms of co-occurrences of objects or scales of objects. For example, burgers, root beers, and buses could appear on the table simultaneously, and their scales would be vastly different from our normal perception. We believe there is a strong semantic correlation between foreground instances and backgrounds in the same image. Therefore, it is reasonable that an instance appears multiple times in the same image. For example, root beers are more inclined to appear in the diet scene, and the increase of root beer instances does not destroy the semantic information in the original image. Accordingly, at the instance-level, we propose Copy-Move data augmentation to copy instances and then move them to other locations in the same image. In addition, we dynamically calculate copy times for each instance according to the rarity of the class to obtain better detection results for the rare class.

For each ground-truth instance g, we can get its class c_g. In general, a long-tailed object detection dataset, such as LVIS [2], will provide the number of images in which the class is annotated. This can implicitly reflect the rarity of each class. Thus, we can use the approach in RFS [2] to define the rarity of the class r_{c_g}:

$$r_{c_g} = \max(1, \sqrt{\frac{t}{f_{c_g}}}) , \tag{1}$$

where f_{c_g} is the fraction of images in which the class is annotated, t is a hyper-parameter. Instances of the rare class will get a larger r_{c_g}. Then, the copy times t_g of the ground-truth instance g can be calculated as:

$$t_g = \begin{cases} \rho & r_{c_g} \geqq r_t \\ 0 & r_{c_g} < r_t \end{cases} , \tag{2}$$

where r_t is the threshold for determining whether the class is rare or not and ρ is the copy times of the rare class. we set $r_t = 3$ in LVIS. Thus, instances of the rare class will be copied more times.

For each image, we use Eq. (2) to calculate the copy times of each instance. If an instance belongs to the rare class and needs to be copied, we perform Copy-Move data augmentation cyclically. Specifically, a scaling factor is randomly selected from [0.8, 1.2] to perform scale jittering on the mask of the instance. Scale jittering randomly is essential to enhance the diversity of the instance. Then select a point from the image randomly as the upper left corner of the placement location, not restricting the scaled instance boundary to exceed the image boundary. Finally, the mask of the original instance is copied and moved to the target location. When all instances have been copied, we first deal with the occlusion between the copied instances. Our approach is that the instance copied first will be occluded by the instance copied later if there is overlap between the two instances. Therefore, we should not set too large copy times. Otherwise, excessive occlusion will destroy the semantic information of the image. Then, we handle the occlusion of the original instances by the copied instance. If the original instance is occluded and reduced by more than 10 pixels in width or height, we filter out the original instance. By Copy-Move data augmentation, the number of instances of the frequent class is maintained while the number of instances of the rare class is increased within a reasonable range, which facilitates long-tailed object detection.

While our main experimental results use the copy times t_g definition above, its precise rule is not crucial. In Appendix 1, we consider other instantiations of the copy times and demonstrate that these can be equally effective.

The previous data augmentations, such as mixup [15] and Copy-Paste [11], did not consider the rarity of the class and directly mixed instances from two images. However, instances of the rare class and instances of the frequent class often appear together, i.e., label co-occurrence. Blindly increasing instances of all classes do not alleviate the class imbalance, as shown in Fig. 1(b). Moreover, it is not efficient to blindly increase the number of instances of all classes. As we will introduce in Section 4.4, the data validity of the previous data augmentations is low. In contrast, our proposed Copy-Move approach introduces the long-tailed information in the dataset into the data augmentation and only increases the number of instances of the rare class. This can alleviate the class imbalance and increase the diversity of rare class instances while having higher data validity. In addition, our approach effectively enhances rare class instances and can be easily embedded in existing long-tailed object detection processing.

3.2 Anchor-Level: Long-Tailed Training Sample Selection

When training an object detector, all ground-truth instances must select their corresponding positive samples. These positive samples are further used for classification and box regression. Thus, positive samples are the ultimate supervision to guide neural network learning. After obtaining more instances of the rare class at the instance-level using the above Copy-Move data augmentation, we propose a Long-Tailed Training Sample Selection (LTTSS) to generate more suitable positive samples for the instances of the rare class. We introduce information about the long-tailed distribution into the training sample selection. Compared to generating positive samples using fixed rules for all classes, our LTTSS approach automatically divides positive samples according to the rarity of the class.

We use the rarity of the class r_{c_g} defined in Eq. (1). Then we can define the mapping between long-tailed scaling factor s_{c_g} and the rarity of the class r_{c_g} as:

$$s_{c_g} = 1 + \frac{\varepsilon}{1 + \mathrm{e}^{-\gamma(r_{c_g} - \mu)}}\,, \tag{3}$$

where ε, γ and μ are hyperparameters and set $\varepsilon = 1$. In this way, the rare class will receive a larger s_{c_g} but no more than 2.

In training sample selection, we first obtain candidate positive samples for each ground-truth instance g. Specifically, for each pyramid level, we select the k_{c_g} anchors closest to the center of ground-truth instance box b_g according to the Euclidean distance. We define $k_{c_g} = \lfloor k \times s_{c_g} \rfloor$ where k is a hyperparameter with a default value of 9. Therefore, the rare class will get more candidate positive samples, at most $2k$. Assuming that there are n_l pyramid levels, a total of $n_l \times k_{c_g}$ candidate positive samples will be obtained for each ground-truth instance and we define it as $\mathcal{C}_g$. Then we calculate intersection of union (IoU) between ground-truth instance box b_g and candidate positive samples $\mathcal{C}_g$ as $\mathcal{I}_g$. Mean m_g and standard deviation v_g of $\mathcal{I}_g$ are then calculated in order to filter. We define the the long-tailed filtering threshold f_t as:

$$f_t = \frac{m_g + v_g}{s_{c_g}}\,. \tag{4}$$

Obviously, the threshold of the rare class is lowered as a way to retain more positive samples. Finally, we select positive candidate samples with IoU greater than or equal to the long-tailed filtering threshold f_t as the final positive samples. In particular, it is necessary to restrict the center of final positive samples inside the ground-truth instance box. The overall flow of LTTSS is shown in Algorithm 1. The bolded pseudo code indicates that information about the long-tailed distribution is used.

Compared with ATSS [31], our algorithm also automatically divides positive samples according to the statistical characteristics of instances. However, instead of using the same rule for all classes, we generate more positive samples for the rare class according to the rarity of the class. Particularly, the increase of positive samples is not blind. Due to the restriction that the center of the positive samples must be located in the center of the ground-truth instance box, we select as many training positive samples as possible for the rare class within a reasonable range.

Algorithm 1: Long-Tailed Training Sample Selection

```
Input:
    $b_g$: a ground-truth instance box;
    $n_l$: the number of pyramid levels;
    $\mathcal{A}_i$: the set of anchors in pyramid level $i$;
    $\mathcal{A}$: the set of anchors in all pyramid levels;
    $r_{c_g}$: the rarity of the class $c_g$;
    $k$: a hyperparameter with a default value of 9
Output:
    $\mathcal{P}_g$: the set of positive samples;
    $\mathcal{N}_g$: the set of negative samples;
compute the long-tailed scaling factor $s_{c_g}$: $s_{c_g} = 1 + \frac{1}{1+\mathrm{e}^{-\gamma(r_{c_g}-\mu)}}$;
compute the number of candidate positive samples $k_{c_g}$ selected
  in each pyramid level: $k_{c_g} = \lfloor k \times s_{c_g} \rfloor$;
for $i$ in $[1, n_l]$ do
    $\mathcal{C}_g = \mathcal{C}_g \cup$ $k_{c_g}$ anchors that closest to the center of $b_g$ according to the
      Euclidean distance in $\mathcal{A}_i$
end
compute IoU $\mathcal{I}_g$ between $b_g$ and $\mathcal{C}_g$; compute mean $m_g$ and standard
  deviation $v_g$ of $\mathcal{I}_g$;
compute the long-tailed filtering threshold $f_t$: $f_t = \frac{m_g + v_g}{s_{c_g}}$;
for each candidate $c$ in $\mathcal{C}_g$ do
    if center of $c$ in $b_g$ and $\mathrm{IoU}(c, b_g) > f_t$ then
        $\mathcal{P}_g = \mathcal{P}_g \cup c$
    end
end
$\mathcal{N}_g = \mathcal{A} - \mathcal{P}_g$;
return $\mathcal{P}_g$, $\mathcal{N}_g$
```

4 Experiments

4.1 Experimental Settings

Dataset. We perform experiments on the challenging LVIS v1 dataset [2]. LVIS is a large vocabulary benchmark for long-tailed object detection, which contains 1203 classes. It provides precise bounding box for various classes with long-tailed distribution. We train our models on the train set, which contains about 100k images. According to the number of images that each class appears in the train split, the classes are divided into three groups: rare (1-10 images), common (11-100 images) and frequent (>100 images). We report results on the val set of 20k images.

Evaluation Metric. We use the widely-used metric AP across IoU threshold from 0.5 to 0.95 to evaluate object detection results. In addition, we also report AP_r, AP_c, AP_f for rare, common and frequent classes to well characterize the long-tailed class performance. Unlike the COCO evaluation process, detection results of classes not listed in the image level labels will not be evaluated.

Implementation Details. We use the same training framework as EFL [9] as our baseline settings. Specifically, we adopt the ResNet-50 [18] initialized by ImageNet [32] pre-trained models as the backbone and feature pyramid network (FPN) [19] as the neck. Besides, we also perform experiments with ResNet-101, a larger backbone to validate the effectiveness of our method. Following the convention, we adopt multi-scale with horizontally flip augmentation during training. Specifically, we randomly resize the shorter edge of the image within $\{640, 672, 704, 736, 768, 800\}$ pixels and keep the longer edge smaller than 1333 pixels without changing the aspect ratio. During the inference phase, we resize the shorter edge of the input image to 800 pixels and keep the longer edge smaller than 1333 pixels without changing the aspect ratio. Our model is optimized by stochastic gradient descent (SGD) with momentum 0.9 and weight decay 0.0001 for 24 epochs. As mentioned in [23], in the one-stage detector, the prior bias of the last layer in the classification branch should be initialized to $-\log\frac{1-\pi}{\pi}$ with $\pi = 0.001$. To avoid abnormal gradients and stabilize the training process, we utilize the gradient clipping with a maximum normalized value of 35. Unlike the EFL settings, we use a total batch size of 8 on 8 GPUs (1 image per GPU) and set the initial learning rate to 0.01 with 1 epoch' warm up. The learning rate decays to 0.001, 0.0001 at the end of epoch 16 and 22, respectively. In addition, we keep the top 300 bounding boxes as prediction results and reduce the threshold of prediction score from 0.05 to 0.0 following [2]. We train all models with RFS [2].

For our proposed Copy-Move, the hyperparameter t used to calculate the rarity of the class is set to 0.001, and the threshold r_t used to determine whether Copy-Move should be used is set to 3. The copy times ρ of the rare class is set to 4, and more details about the impact of this hyperparameter are showcased in Sect. 4.3. In particular, we perform Copy-Move data augmentation with probability 0.5 and close it for the last 3 epochs, when the learning rate is decayed for the last time.

For our proposed LTTSS, we tile two anchors per location on the image with the anchor scale $\{6, 8\}$ with $k = 18$ to cover more potential candidates. In Eq. (3), hyperparameter γ and μ can adjust the slope and central region of the curve. They work together to control the influence range of long-tailed scaling factor s_{c_g}. We set $\gamma = 5.0$ and $\mu = 2.5$. More details about the impact of γ and μ are showcased in Sect. 4.3.

4.2 Benchmark Results

Table 1 demonstrates the effectiveness of our proposed REDet. We compare our approach with other works that report state-of-the-art performance and other augmentations that can significantly boost object detection performance. With ResNet-50 backbone, our proposed REDet achieves an overall 28.3% AP, which improves the baseline by 0.8% AP, and even achieves 1.4 points improvement on the rare class. It can be seen that Copy-Move and LTTSS work collaboratively to realize the instances equilibrium and positive samples equilibrium in long-tailed object detection and dramatically improve the performance of the rare class without sacrificing the frequent class. Compared with other state-of-the-art methods like EQL [6], EQLv2 [8], BAGS [13] and Seesaw Loss [7], our proposed method outperforms them by 3.2% AP, 2.8% AP, 2.3% AP and 1.9% AP, respectively. In addition, we also add data augmentations such as mixup [15], InstaBoost [16], Copy-Paste [11] to the baseline. In particular, Copy-Paste was trained in [11] in a decoupled strategy. Specifically, in the first stage, they trained the object detector for 180k steps using a 256 batch size. Then they fine-tuned the model with 36 epochs in the second stage. For a fair comparison, we train Copy-Paste with large scale jittering on the image size of 1024×1024 in an end-to-end strategy for 24 epochs. More training details can be found in Appendix 2. Compared to these methods, our proposed method outperforms them by 3.0% AP, 1.4% AP and 1.1% AP, respectively. This is mainly because the existing data augmentations do not consider the information of the long-tailed distribution in the dataset. Mixup and Copy-Paste simply mix or paste the instances from two images together. They do not consider the label co-occurrence problem, which exacerbates the imbalance and ultimately yields lower detection results. InstaBoost jitter the location of instances in all classes by calculating the location probability map, which hardly changes the number of instances in each class. It shows that blindly increasing the number of instances or jittering instances in long-tailed object detection is inefficient, requiring significant computational resources while failing to improve the final performance. Our proposed REDet introduces class rarity in data augmentation and positive sample sampling, which alleviates the imbalance in the dataset and leads to better performance in AP and AP_r.

We conduct experiments with larger ResNet-101 backbone. Our approach can still obtain consistent improvements in overall AP and AP_r by 1.0% and 1.4%, respectively. Compared to other data augmentations, our proposed approach still performs better in long-tailed object detection. It indicates that our REDet can alleviate the imbalance across different backbones. Our method achieves 30.2% AP and establishes a new state-of-the-art.

4.3 Ablation Study

We conduct a series of comprehensive ablation studies to verify the effectiveness of the proposed REDet. For all experiments, we use ResNet-50 as backbone for 24 epochs.

Table 1: Comparison with other state-of-the-art approaches and other augmentations on LVIS v1 val set. † indicates that the reported result is directly copied from referenced paper. + indicates that the augmentation is added to the baseline.

backbone	method	strategy	AP	AP$_r$	AP$_c$	AP$_f$
ResNet-50	*other methods*					
	EQL† [6]	end-to-end	25.1	15.7	24.4	30.1
	EQLv2† [8]	end-to-end	25.5	16.4	23.9	31.2
	BAGS† [13]	decoupled	26.0	17.2	24.9	31.1
	Seesaw Loss† [7]	end-to-end	26.4	17.5	25.3	31.5
	EFL (Baseline)† [9]	end-to-end	27.5	20.2	26.1	32.4
	augmentations					
	+ mixup [15]	end-to-end	25.3	18.5	23.3	30.5
	+ InstaBoost [16]	end-to-end	26.9	19.7	25.6	31.5
	+ Copy-Paste [11]	end-to-end	27.2	21.3	25.8	31.5
	REDet (Ours)	end-to-end	**28.3**	**21.6**	**26.8**	**32.9**
ResNet-101	*other methods*					
	EQLv2† [8]	end-to-end	26.9	18.2	25.4	32.4
	BAGS† [13]	decoupled	27.6	18.7	26.5	32.6
	Seesaw Loss† [7]	end-to-end	27.8	18.7	27.0	32.8
	EFL (Baseline)† [9]	end-to-end	29.2	23.5	27.4	33.8
	augmentations					
	+ mixup [15]	end-to-end	28.8	21.4	27.1	33.8
	+ InstaBoost [16]	end-to-end	28.6	22.0	27.2	33.2
	+ Copy-Paste [11]	end-to-end	29.7	24.1	27.8	**34.4**
	REDet (Ours)	end-to-end	**30.2**	**24.9**	**28.5**	34.3

Influence of Components in Our Approach. There are two components in our REDet, Copy-Move and LTTSS. As shown in Table 2, both Copy-Move and LTTSS play significant roles in our approach. Copy-Move can achieve an improvement from 27.5% AP to 28.1% AP, and achieve 0.6 points improvement on the rare class without degrading the performance of the frequent class. Our approach calculates the copy times of an instance based on the rarity of the class, which alleviates the problem of the lack of rare class instances and makes the dataset more balanced. LTTSS generates more supervision for the rare class within a reasonable range at positive sample sampling and achieves an improvement from 27.5% AP to 27.8% AP. Combining the two components, our REDet takes the performance of the baseline from 27.5% to 28.3%. In particular, we can achieve a 1.4% improvement in the rare class. This is due to the fact that the instances and positive samples of the rare class grow within a reasonable range at the same time.

Table 2: Ablation study of each component in our approach. Copy-Move and LTTSS indicate Copy-Move augmentation, and long-tailed training sample selection, respectively.

Copy-Move	LTTSS	AP	AP_r	AP_c	AP_f
		27.5	20.2	26.1	32.4
✓		28.1	20.8	26.7	32.8
	✓	27.8	20.4	26.3	32.8
✓	✓	**28.3**	**21.6**	**26.8**	**32.9**

Influence of Number of Instances and Positive Samples. It can be seen that our proposed approaches, Copy-Move and LTTSS, increase the number of instances and positive samples of the rare class, respectively, according to the rarity of the class. In particular, the beneficial performance is brought by our specially designed rules that exploit the long-tailed information in the dataset rather than by brutally increasing the number of instances and positive samples. To prove this, we randomly and uniformly increase the number of instances and positive samples for all classes until they are close to the number of Copy-Move and LTTSS, respectively. Table 3 shows the experimental results. Compared to baseline, rand Copy-Move copies 3 instances for each class. This approach can not alleviate the degree of imbalance and can hardly have an impact on AP. The performance of the various classes was virtually unchanged. On the other hand, we reduce the threshold for all classes in positive sample sampling instead of calculating based on the rarity of the class. It can be seen that compared to baseline, the random LTTSS even reduces 0.1% AP and 1.0% AP_r due to blindly lowering the threshold for all classes.

Table 3: Ablation study of the number of instances and positive samples. Rand Copy-Move and rand LTTSS indicate randomly and uniformly increasing the number of instances and positive samples for all classes, respectively.

method	number	AP	AP_r	AP_c	AP_f
	instances				
Baseline	429.8k	27.5	20.2	26.1	32.4
rand Copy-Move	432.9k	27.5	20.2	25.9	32.5
Copy-Move	433.3k	**28.1**	**20.8**	**26.7**	**32.8**
	positive samples				
Baseline	24.15M	27.5	20.2	26.1	32.4
rand LTTSS	24.38M	27.4	19.2	26.0	32.6
LTTSS	24.38M	**27.8**	**20.4**	**26.3**	**32.8**

Influence of the Hyperparameter. We study the hyperparameters, i.e., ρ, γ, μ, adopted in different components of our REDet. In Table 4(a), we explore ρ in Copy-Move. ρ controls the times that the rare class instance is copied. When ρ is too small, the number of rare instances is still insufficient to alleviate the imbalance of the dataset. When ρ is too large, a more serious occlusion occurs between the copied and original instances during the movement. This corrupts the semantic information in the image and makes object detection more difficult. We find that $\rho = 4$ achieves the best performance. In Table 4(b), we explore γ, μ in LTTSS. γ and μ control the slope and central region of the curve and further control the influence range of the long-tailed scaling factor s_{c_g} as shown in Fig. 3. Results show that $\gamma = 5.0, \mu = 2.5$ achieves the best performance.

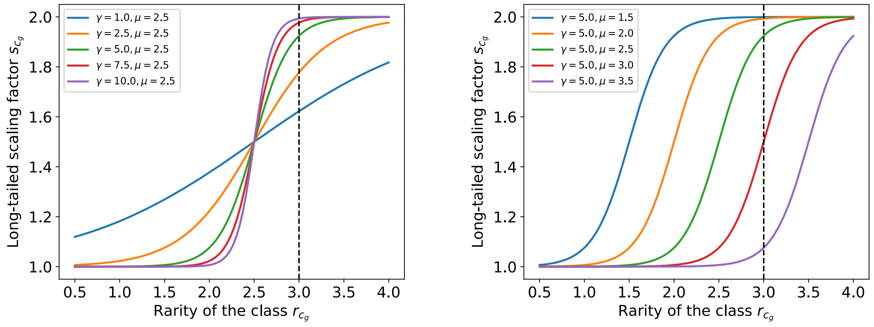
Table 4: Ablation study of the hyperparameter ρ, γ and μ, $\rho = 4, \gamma = 5.0, \mu = 2.5$ is adopted as the default setting in other experiments. (a) hyperparameter ρ. (b) hyperparameters γ and μ.

(a)

ρ	AP	AP_r	AP_c	AP_f
1	27.7	20.0	26.2	32.7
2	27.9	20.1	26.5	32.8
3	27.9	20.3	26.4	**32.9**
4	**28.1**	**20.8**	**26.7**	32.8
5	27.7	19.1	26.5	32.8
6	27.5	18.3	26.4	32.8

(b)

γ	μ	AP	AP_r	AP_c	AP_f
5.0	2.0	27.7	20.2	26.1	32.8
5.0	2.5	27.8	**20.4**	26.3	32.8
5.0	3.0	27.8	20.0	26.3	32.8
3.0	2.5	27.6	19.9	26.1	32.7
10.0	2.5	27.5	19.4	26.1	32.7
15.0	2.5	27.3	18.6	25.7	32.8



(a) (b)

Fig. 3: Comparison of the long-tailed scaling factor s_{c_g} with different hyperparameters. The black vertical line represents the demarcation line between the rare and other classes. (a) Different γ with $\mu = 2.5$. (b) Different μ with $\gamma = 5.0$.

4.4 Data Validity Analysis

Performing data augmentation uniformly for all classes is inefficient. We propose a metric called data validity to measure this. We quantitatively demonstrated the data validity of several data augmentations and our proposed Copy-Move. In detail, We define data validity v_d as $v_d = \frac{\triangle \mathrm{AP}}{\triangle n_i}$, where $\triangle n_i$ denotes increase in the number of instances and $\triangle$AP denotes increase in AP. Specifically, we count the number of all instances augmented during 1 epoch while recording the final performance gain for each method. Table 5 shows the detailed results. Mixup and Copy-Paste blindly increase the number of instances by about 57%, 33%, respectively. However, increasing the number of instances of all classes does not alleviate the data imbalance and introduces severe occlusion. Their data validity is only -8.9e-6 and -2.1e-6, respectively. InstaBoost jitters instance's position through the probability map, which may lead to missing instances. Despite the positive data validity 6.5e-5, it does not lead to performance gains. Our approach increases the instances of the rare class according to the rarity and maintains the frequent class instances unchanged, which ultimately yields greater data validity 1.7e-4.

Table 5: Data validity v_d of several data augmentations and our proposed Copy-Move. $\triangle$ indicates the increase compared to baseline.

method	n_i	$\triangle n_i$	AP	$\triangle$AP	v_d
Baseline [9]	429.8k	0.0k	27.5	0	
mixup [15]	676.8k	247.0k	25.3	-2.2	-8.9e-6
InstaBoost [16]	420.6k	-9.2k	26.9	-0.6	6.5e-5
Copy-Paste [11]	573.4k	143.6k	27.2	-0.3	-2.1e-6
Copy-Move (Ours)	433.3k	3.5k	28.1	+0.6	**1.7e-4**

5 Conclusion

In this paper, we boost dense long-tailed object detection from a new data-centric view. A rare data centric detection framework REDet is proposed to alleviate data imbalance while increasing training data. Novel Copy-Move data augmentation and Long-Tailed Training Sample Selection (LTTSS) work together to dynamically increase the number of instances and positive samples according to the rarity. Our proposed approach is the first to bridge the gap of handling long-tailed distribution data at the instance-level and anchor-level. It brings significant improvements with notably boosting on the rare class. Combining the two components, our REDet beats existing state-of-the-art approaches on the challenging LVIS v1 benchmark, which shows the superiority of our method. We hope that our REDet could be a standard procedure when training one-stage long-tailed object detection models.

Acknowledgments. This work was supported by the National Key Research and Development Program of China (Grant No. 2018YFE0203801).

References

1. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. arXiv preprint arXiv:2110.04596 (2021)
2. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2019)
3. Wu, J., Song, L., Wang, T., Zhang, Q., Yuan, J. In: Forest R-CNN: Large-Vocabulary Long-Tailed Object Detection and Instance Segmentation. Association for Computing Machinery, New York, NY, USA (2020) 1570–1578
4. Wang, T., Li, Y., Kang, B., Li, J., Liew, J., Tang, S., Hoi, S., Feng, J.: The devil is in classification: A simple framework for long-tail instance segmentation. In Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., eds.: Computer Vision – ECCV 2020, Springer International Publishing (2020) 728–744
5. Feng, C., Zhong, Y., Huang, W.: Exploring classification equilibrium in long-tailed object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). (2021) 3417–3426
6. Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., Yan, J.: Equalization loss for long-tailed object recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2020)
7. Wang, J., Zhang, W., Zang, Y., Cao, Y., Pang, J., Gong, T., Chen, K., Liu, Z., Loy, C.C., Lin, D.: Seesaw loss for long-tailed instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2021) 9695–9704
8. Tan, J., Lu, X., Zhang, G., Yin, C., Li, Q.: 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). (2021) 1685–1694
9. Li, B., Yao, Y., Tan, J., Zhang, G., Yu, F., Lu, J., Luo, Y.: Equalized focal loss for dense long-tailed object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2022) 6990–6999
10. Hu, X., Jiang, Y., Tang, K., Chen, J., Miao, C., Zhang, H.: Learning to segment the tail. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2020)
11. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: 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). (2021) 2918–2928
12. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. In: International Conference on Learning Representations. (2020)
13. Li, Y., Wang, T., Kang, B., Tang, S., Wang, C., Li, J., Feng, J.: 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 (CVPR). (2020)
14. Rogers, A.: Changing the world by changing the data. In: ACL. (2021)
15. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations. (2018)
16. Fang, H.S., Sun, J., Wang, R., Gou, M., Li, Y.L., Lu, C.: Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). (2019)

17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., Weinberger, K., eds.: Advances in Neural Information Processing Systems. Volume 25., Curran Associates, Inc. (2012)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
19. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
20. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2015)
21. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., eds.: Advances in Neural Information Processing Systems. Volume 28., Curran Associates, Inc. (2015)
22. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018)
23. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2017)
24. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In Leibe, B., Matas, J., Sebe, N., Welling, M., eds.: Computer Vision – ECCV 2016, Cham, Springer International Publishing (2016) 21–37
25. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
26. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: Centernet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). (2019)
27. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). (2019)
28. Alexey, B., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
29. Zhong, Z., Cui, J., Liu, S., Jia, J.: Improving calibration for long-tailed recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2021) 16489–16498
30. Zang, Y., Huang, C., Loy, C.C.: Fasa: Feature augmentation and sampling adaptation for long-tailed instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). (2021) 3457–3466
31. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2020)
32. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. (2009) 248–255