

OOD-CV Challenge Report

October 10, 2022

1 Team details

- Challenge track: Detection Track
- Team name: detectors_218
- Team leader name: Zining Chen
- Team leader address, phone number, and email:
Address: Beijing University of Posts and Telecommunications, Beijing, China
Phone number: +86-15643117621
Email: chenzn@bupt.edu.cn
- Rest of the team members: Tianyi Wang
- Team website URL: None
- Affiliation: Beijing University of Posts and Telecommunications
- User names on the OOD-CV Codalab competitions: detectors_218

- Link to the codes of the solution(s):
<https://github.com/VincentWangty/oodcv>

2 Contribution details

- Title of the contribution:
 OCP: An Effective Data Augmentation Method with Two-stage Framework for Out-of-Distribution Object Detection
- General method description:
 In this paper, we propose a simple and effective two-stage framework with Object-based CopyPaste (OCP) for out-of-distribution object detection. Also, we design Nuisance-Specific Weighted Boxes Fusion (NS-WBF), a rank-based post-processing method. Firstly, OOD-CV challenge aims to train a well-generalized model on both train images and unknown-distributed target nuisances, where occlusion accounts for a fairly large proportion. Thus we propose OCP to tackle occlusion as the key challenge. Secondly, as test images in phase 2 can be leveraged to finetune models in phase 1, we propose an universal two-stage framework for out-of-distribution (OOD) object detection. We train the object detector using images with ground truth bounding boxes, where only basic data augmentation methods are used to increase the diversity of data. Then we finetune models using high-quality pseudo-labels of test images in phase 2 with OCP data augmentation method. Finally, we design NS-WBF to fuse diverse models on different nuisances according to their ranks. Under this circumstances, our framework can successfully handle all nuisances, especially occlusion, and fully utilize domain information to increase the robustness and generalization ability of models.
- Description of the particularities of the solutions deployed for each of the tracks:
 Our solution mainly consists of three particularities on object detection track.
 1. Two-stage Framework
 First, input independent and identically distributed (IID) images and apply basic data augmentation methods, including Resize, Flip, Mixup [23], Cutout [5] and PhotoMetricDistortion (PMD), to get

IID-Aug images for stage-1 training. Second, inference OOD images of phase 2 to generate high-quality pseudo-labels based on the trained detectors, and different filter mechanisms, confidence score or Intersection of Union (IoU) of bounding boxes, are utilized to filter inaccurate and redundant boxes. Finally we adopt the former with the threshold of 0.7 based on quantitative criterion, which is the estimation on the number of bounding boxes. Last, finetune the trained detectors using dataset-specific augmentation method (OCP for ROBIN dataset [24], details are demonstrated as follows).

2. Object-based CopyPaste (OCP)

OCP is specifically designed for occluded objects without segmentation labels, which copy objects by ground truth bounding boxes and paste them to images is fairly appropriate for a synthetic occlusion-based dataset. However, diverse hyper-parameters, such as the number of pasted objects, IoU between pasted objects and objects in current image and the position of the pasted objects, will affect the representation learning of models. To reduce the gap between the distribution of IID and OOD images to the utmost extent, we select one pasted objects per image, the position of the pasted object is the exact coordinates of its original image and padding is added if the size of pasted objects is larger.

As the training process follows a two-stage manner, for better representation learning of foreground objects in stage-1, OCP is only used in stage-2. Specifically, images and ground truth bounding boxes in stage-2 are test images and pseudo labels generated by stage-1 detectors, respectively. Therefore, copy objects in training images and paste them on test images can increase the diversity of data and avoid over-fitting on pseudo labels.

3. Nuisance-Specific Weighted Boxes Fusion (NS-WBF)

NS-WBF is stemmed from Weighted Boxes Fusion (WBF) [18], using all bounding boxes to get weighted sum coordinates of bounding boxes. We further design NS-WBF to fuse diverse models on different nuisances according to their ranks. Detailed algorithm and equation are demonstrated in Section 4.

• References:

- [1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms—improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017.
- [2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. *computer vision and pattern recognition*, 2017.
- [3] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Grid-mask data augmentation. *arXiv preprint arXiv:2001.04086*, 2020.
- [4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. *international conference on computer vision*, 2017.
- [5] Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv: Computer Vision and Pattern Recognition*, 2017.
- [6] 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. *computer vision and pattern recognition*, 2021.
- [7] Yuqi Gong, Xuehui Yu, Yao Ding, Xiaoke Peng, Jian Zhao, and Zhenjun Han. Effective fusion factor in fpn for tiny object detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1160–1168, 2021.
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv: Computer Vision and Pattern Recognition*, 2015.
- [9] Pavel Izmailov, D. A. Podoprikin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *uncertainty in artificial intelligence*, 2018.
- [10] Tingting Liang, Xiaojie Chu, Yudong Liu, Yongtao Wang, Zhi Tang, Wei Chu, Jingdong Chen, and Haibing Ling. Cbnetv2:

A composite backbone network architecture for object detection. *arXiv preprint arXiv:2107.00420*, 2021.

- [11] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. *arXiv: Computer Vision and Pattern Recognition*, 2016.
- [12] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. *computer vision and pattern recognition*, 2018.
- [13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *international conference on computer vision*, 2021.
- [14] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th International Conference on Pattern Recognition (ICPR’06)*, volume 3, pages 850–855. IEEE, 2006.
- [15] Siyuan Qiao, Liang-Chieh Chen, and Alan L. Yuille. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. *computer vision and pattern recognition*, 2021.
- [16] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. *computer vision and pattern recognition*, 2020.
- [17] Hamid Rezaatofghi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. *computer vision and pattern recognition*, 2019.
- [18] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, 2021.
- [19] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

- [20] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022.
- [21] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *computer vision and pattern recognition*, 2016.
- [22] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander J. Smola. Resnest: Split-attention networks. *computer vision and pattern recognition*, 2020.
- [23] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.
- [24] Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. Ood-cv: A benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.
- [25] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12993–13000, 2020.

- Representative image / diagram of the method(s):
Please refer to Figure 1.

3 Global Method Description

[* Indicates method used in competition test results.]

- Total method complexity: Details are shown in Table 1
- Model Parameters: Details are shown in Table 1
- Run Time: Details are shown in Table 1

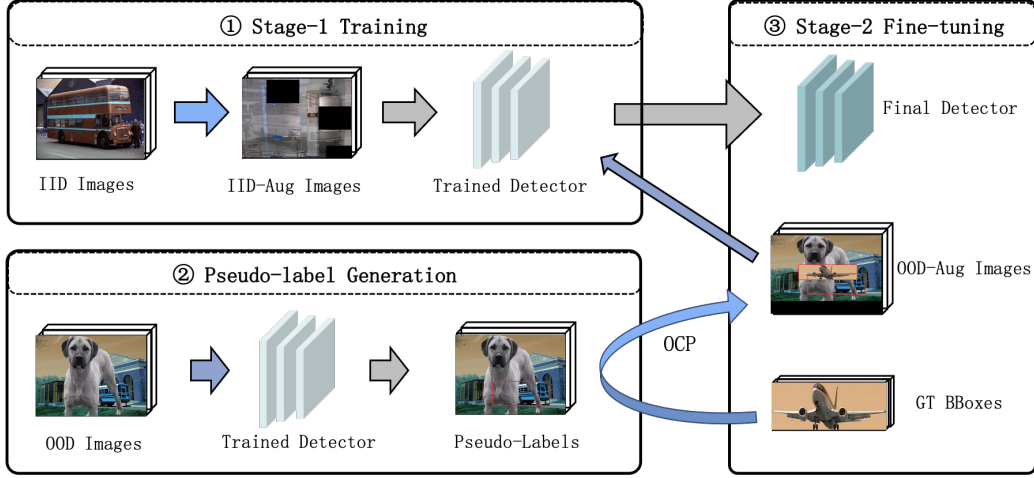


Figure 1: **Our two-stage framework with Object-based CopyPaste (OCP) for out-of-distribution object detection.** It consists of three stages. ① **Stage-1 training:** Use IID-Aug images for stage-1 training. ② **Pseudo-labeling Generation:** Inference OOD images of phase 2 to generate high-quality pseudo-labels based on the trained detectors. ③ **Stage-2 Fine-tuning:** Finetune the trained detectors with OCP augmentation method.

- Which pre-trained or external methods / models have been used:
 Backbone: ResNet-50 [8], ResNet-101 [8], ResNeXt-101-64x4d [21], ResNest-101 [22], Regnetx-12GF [16]
 Neck: Feature Pyramid Network [11] (FPN)
 Head: Cascade R-CNN [2], Yolov7 [20], Detectors[15]
 Data Augmentations: Flip, Cutout, Mixup, PhotoMetricDistortion
- Training description:
 We totally adopt 6 backbones, 1 neck and 3 heads to get final results. Detailed information is shown in Table 2.
 Here we expound our original baseline on ResNet-50, FPN and Cascade R-CNN for stage-1 training, where first we add Deformable Convolution Network [4] (DCN) structure to keep spatial-invariant. Besides, we select GIoU Loss [17] for regressing bounding boxes and we adjust hyper-parameters on IoU threshold in assigner. Furthermore, we adopt Mixup, Resize, Flip, Cutout and PMD augmentations to increase the

Model	Size	FLOPs	Parameters	Run time
Cascade R-CNN ResNet-50	(500, 375)	72.28G	69.54M	5.4h
Cascade R-CNN ResNet-101	(500, 375)	79.73G	89.23M	7h
Cascade R-CNN ResNeXt-101-64x4d	(1333, 800)	492.01G	131.82M	10h
Cascade R-CNN ResNest-101	(1000, 600)	573.24G	95.46M	8.4h
Cascade R-CNN Regnetx-12GF	(1333, 800)	413.85G	88.91M	7.8h
Detectors ResNet-101	(500, 375)	78.92G	188.57M	10h
Yolov7-w6	(640, 640)	102.8G	81.1M	6.4h
Average		258.96G	106.35M	7.86h

Table 1: Detailed information on total method complexity, model parameters and run time of different models.

diversity of data. Then for stage-2 finetuning, we add OCP augmentation with other structures unchanged. We train stage-1 model using SGD optimizer with step scheduler, learning rate is initialized to $2e^{-2}$ for 72 epochs and scale it by 0.1 after 56 epochs and 64 epochs, weight decay is $1e^{-4}$. For stage-2, we finetune the trained models for extra 12 epochs, learning rate is $2e^{-3}$ and scale it by 0.1 after 8 epochs, and finally use Stochastic Weight Averaging[9] (SWA) strategy to increase model generalization. Batch size is 4 or 8 according to input size and multi-scale training is alternatively applied.

- Testing description:
Test-Time Augmentation (TTA) is applied to enhance images and different TTA combinations are conducted to get the best strategy, including flip with probability of 0.5, the same scale in training and PMD. Besides, we adopt Soft-Non-Maximum-Suppression [1](Soft-NMS) to remove redundant bounding boxes for single model, which we conduct ablation studies for an optimal confidence score, 0.005, while 0.05 drops 0.72% and 0.001 performs worse when model fusion strategies are implemented. Lastly, we apply NS-WBF to ensemble models, details are demonstrated in Section 4.
- Quantitative and qualitative advantages of the proposed solution:
Our solution on object detection track has the following three advan-

Backbone	Head	Loss	Data Augs	Input size
ResNet-50	Cascade R-CNN	GIoU	Basic+OCP	Multi-scale
ResNet-101	Cascade R-CNN	CIoU [25]	Basic+OCP	(1333,800)
ResNet-101	Detectors	GIoU	Basic+OCP	(500,375)
ResNeXt-101	Cascade R-CNN	CIoU	Basic+OCP	(500,375)
Regnetx-12GF	Cascade R-CNN	CIoU	Basic+OCP	(1333,800)
ResNest-101	Cascade R-CNN	CIoU	Basic+OCP	(1333,800)
ELAN-Net	Yolov7-w6	SmoothL1	Basic-Yolo	(640,640)

Table 2: Detailed training description of all models. Main differences are the selection of backbone, head, loss, data augmentations, input size and other settings remain the same as the above baseline. Basic+OCP denotes using baseline data augmentation settings for two-stage, while Basic-Yolo denotes the original Yolov7-w6 settings. Multi-scale denotes using size between (1333,400) and (1333,800).

tages. Firstly, universal and generalized. Our two-stage framework is universal to all OOD situation, which stage-1 increases the ability of common feature extraction, and pseudo-label usage in stage-2 allows model to learn out-of-distribution samples with specific-designed OCP to fit the distribution of test set. Secondly, simple and easy-plugged. OCP requires no segmentation labels in original CopyPaste [6] augmentation, thus facilitating the use in practical scene. Thirdly, efficient and memory-saving. OCP consumes no extra computing resources and two-stage framework is trained without complex modules.

- Results of the comparison to other approaches (if any) :
Firstly, we clarify the development process in phase 1, as illustrated in Table 3. These methods are added to Baseline step-by-step and significantly improve performance. Besides, we conduct sufficient ablation studies to demonstrate the optimal selection of our method by changing network structure, loss type and data augmentation methods. We attempt to replace FPN structure into state-of-the-art neck structure, PAFPN [12] and fusion factor [7], but they decrease the performance by 1.13% and 0.85%. Smooth-L1 and Label Smoothing CE Loss [19] still worsen the performance by 1.06% and 2.72%. Random Gray and

Methods	mAP
Baseline	57.83%
+Cutout	63.97%
+Mixup	66.54%
+CIoU	67.60%
+Hyper-parameter Adjustment	69.54%
+Two-stage Framework	71.53%
+OCP	72.97%

Table 3: Results of the development of baseline model in phase 1. Baseline indicates ResNeXt-101 backbone network and + denotes adding the method based on the previous experimental settings. Hyper-parameter adjustment includes the number of holes in Cutout and IoU threshold in assigner. OCP in phase 1 are implemented on original train images and ground truth bounding boxes.

Gaussian Noise also drops 0.19% and 5.07% respectively.

In order to further prove the effectiveness of OCP, we select state-of-the-art data augmentation methods on occlusion, Gridmask [3] and Cutout. The former decreases the mAP of 2.09%, and the latter achieves approximately the same with OCP, but OCP can be used upon Cutout for further improvement of 1.44%.

- Novelty of the solution and if it has been previously published:
Firstly, to our best knowledge, the two-stage framework is a novel method for OOD object detection. If models are trained in an end-to-end manner, the performance worsens probably because dataset-specific augmentations may impair representation learning of foreground objects. Thus we point out that training models in stage-1 with basic data augmentations, while finetuning models in stage-2 with dataset-specific data augmentations will be an universal solution to all OOD object detection scenarios. Secondly, OCP is designed for occlusion object detection without segmentation labels. During our analysis of ROBIN dataset, we figure out that occlusion serves as the key challenge and most of them are synthetic. Thus we copy objects by ground truth bounding boxes and paste them to other images under several

conditions as mentioned in Section 2. Lastly, NS-WBF is a creative post-processing method specifically designed for this challenge, which weights of different models in different nuisances follow a rank-based manner. These three novelties are proposed through our analysis on object detection track of OOD-CV challenge, which are not previously published.

4 Ensembles and fusion strategies

- Describe in detail the use of ensembles and/or fusion strategies (if any): Based on Weighted Boxes Fusion (WBF), we further improve a rank-based Nuisance-Specific WBF, which diverse models have different weights. We normalize weight to $(0,1]$ for all models on each nuisance and weights are calculated as follows,

$$\begin{aligned} S_n &= [S_{n,1}, \dots, S_{n,k}], \\ O_n &= \text{Argsort}(S_n), \\ W_{n,O_{n,i}} &= (i + 1)/k, \quad i \in [0, k), i \in \mathbb{Z} \end{aligned} \tag{1}$$

where S_n denotes the mAP scores n_{th} nuisances, $S_{n,i}$ denotes the mAP score of i_{th} model in n_{th} nuisances, i indicates the current model index and k indicates the number of models, O_n denotes the indices that sort S_n , $O_{n,i}$ denotes the rank of i_{th} model in n_{th} nuisances and $W_{n,O_{n,i}}$ denotes the weight of i_{th} model in n_{th} nuisances.

- What was the benefit over the single method? : Traditional post-processing methods like NMS [14] and soft-NMS are applied on single model, which only exclude bounding boxes but have no effect on improving accuracy. WBF uses all the bounding boxes with weights to get weighted sum coordinates of bounding boxes, which can effectively improve precision. NS-WBF further uses advantages of diverse models on different nuisances and fuse models on each of the nuisance to guarantee the best use of single model.
- What were the baseline and the fused methods? : In phase 1, baseline denotes the maximum mAP among all the models which achieves 72.97% and WBF, NS-WBF respectively achieves 76.63%, 77.65% for fused method. In phase 2, baseline achieves 63.34% and NS-WBF achieves 65.63%.

5 Technical details

- Language and implementation details (including platform, memory, parallelization requirements) :

Ubuntu version	Ubuntu 18.04.4
CPU	Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz
RAM	502GB
GPU	Tesla V100 SXM2 32GB($\times 8$)
CUDA version	10.2
Programming language	Python3.7
Deep learning framework	Pytorch (torch 1.7.0, torchvision 0.8.0)

Table 4: Environments and requirements

- Human effort required for implementation, training and validation?:
Two team members work for around two months to finish the whole competition, where training and validation consumes a large proportion of time and implementation consumes more time on thinking and understanding of the challenge.
- Training/testing time? Runtime at test per image :
We train models, except Yolov7-w6, for 72 epochs and finetune for 12 epochs, and Yolov7-w6 for 100 epochs and 12 epochs. Details are shown in Table 5
- Comment the efficiency of the proposed solution(s)? :
OCP consumes no extra computing resources and two-stage framework is trained without high-cost modules, thus the efficiency of the overall solution is fairly high, compared with current state-of-the-art models, such as CBNetV2 [10] + Swin-Transformer [13].

6 Other details

- General comments and impressions of the OOD-CV challenge. :
OOD-CV challenge successfully promotes the development of out-of-

Model	Training time	Testing time (per image)
Cascade R-CNN ResNet-50	5.4h	150ms
Cascade R-CNN ResNet-101	7h	190ms
Cascade R-CNN ResNeXt-101-64x4d	10h	370ms
Cascade R-CNN ResNeSt-101	8.4h	300ms
Cascade R-CNN Regnetx-12GF	7.8h	120ms
Detectors ResNet-101	10h	200ms
Yolov7-w6	6.4h	12ms

Table 5: Detailed information on training time and testing time per image.

distribution generalization, one of the frontier fields of computer vision. Meanwhile, as a core downstream task of computer vision, object detection plays a crucial role in practical applications, such as automatic driving and remote sensing communication. Thus, out-of-distribution in object detection is a challenging scenario and also the bottleneck in practical use for long, which worths profound study by worldwide researchers. Also, we are grateful that the presentation of ROBIN dataset can accelerate the growth in OOD generalization. 10 traditional classes from 6 diverse nuisances and an IID subset are really fantastic works, which model performance can be evaluated on different situations. And during phase 1, we find out several misannotations in ROBIN dataset, which will worsen the performance of models, such as chair in diningtable are not annotated, resulting in wrong negative samples for detectors. We hope these reminders can help you revise the annotations of ROBIN to some extent. Last but not least, we really appreciate your work not only on source codes and dataset, but also on evaluation platform, forum and fair rules.

- Other comments: Thank for your quick and clear replies during model development in phase 1!