OOD-CV Challenge Report

September 18, 2023

1 Team details

- Challenge track:
  OOD-CV Challenge 2023 (Detection Track - Self-supervised pretrain leaderboard)
- Team name:
  ll_ly
- Team leader name:
  Wenxuan She
- Team leader address, phone number, and email:
  South Campus of Xidian University, Xi'an, Shaanxi Province, China, +8618706831828, swx_sxpp@qq.com
- Rest of the team members:
  Yu Liu, Yuting Yang, Fang Liu, Xu Liu
- Team website URL:
  None
- Affiliation:
  Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University
- User names on the OOD-CV Codalab competitions:
  ll_ly

- Link to the codes of the solution(s):
  https://github.com/swx3027925806/PaddleDetection-OOD-Det

2 Contribution details

- Title of the contribution :
  Semi-supervised Object Detection Strategy Based on Multi-scale WBF on OOD

- General method description:
  This method is divided into three parts in the training stage, namely: self-monitoring training, supervised training of object detection and semi-supervised training of object detection.

  In the self-monitoring training, the backbone network is selected as ViT model[1], and the self-monitoring algorithm is CAE algorithm[2], and the self-monitoring training task is completed on ImageNet1K.

  In the supervised object detection, the PPYoloE[3] model is selected to train on the training set, and different data enhancement strategies are supplemented for different OOD situations to improve the network performance.

  Thirdly, semi-supervised learning is done by DenseTeacher[4], and its semi-supervised data set is the training set in different tracks of the competition and the test set data in the first stage.

  In the testing stage, the multi-scale WBF is added as the data fusion strategy during testing.

  This method designs different data enhancement strategies for the weak parts of OOD data, which will be displayed in subsequent ablation experiments.

  Finally, the scheme achieved a mAP of 49.4.

- Description of the particularities of the solutions deployed for each of the tracks :
  The deep learning framework adopted in this scheme is PaddlePaddle2.4.0[5], and the relevant environment needs to be configured when the code is reproduced. The adopted code architecture is the modified PaddleDetection, which can be downloaded

directly from the provided github. Start the training automation script by running train.py and start the test script by running test.py

- References:

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[2] Antoine Chevrot, Alexandre Vernotte, and Bruno Legeard, "Cae: Contextual auto-encoder for multivariate time-series anomaly detection in air transportation," *Computers & Security*, vol. 116, pp. 102652, 2022.

[3] Shangliang Xu, Xinxin Wang, Wenyu Lv, Qinyao Chang, Cheng Cui, Kaipeng Deng, Guanzhong Wang, Qingqing Dang, Shengyu Wei, Yuning Du, et al., "Pp-yoloe: An evolved version of yolo," *arXiv preprint arXiv:2203.16250*, 2022.

[4] Hongyu Zhou, Zheng Ge, Songtao Liu, Weixin Mao, Zeming Li, Haiyan Yu, and Jian Sun, "Dense teacher: Dense pseudo-labels for semi-supervised object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 35–50.

[5] Yanjun Ma, Dianhai Yu, Tian Wu, and Haifeng Wang, "Paddlepaddle: An open-source deep learning platform from industrial practice," *Frontiers of Data and Domputing*, vol. 1, no. 1, pp. 105–115, 2019.

[6] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

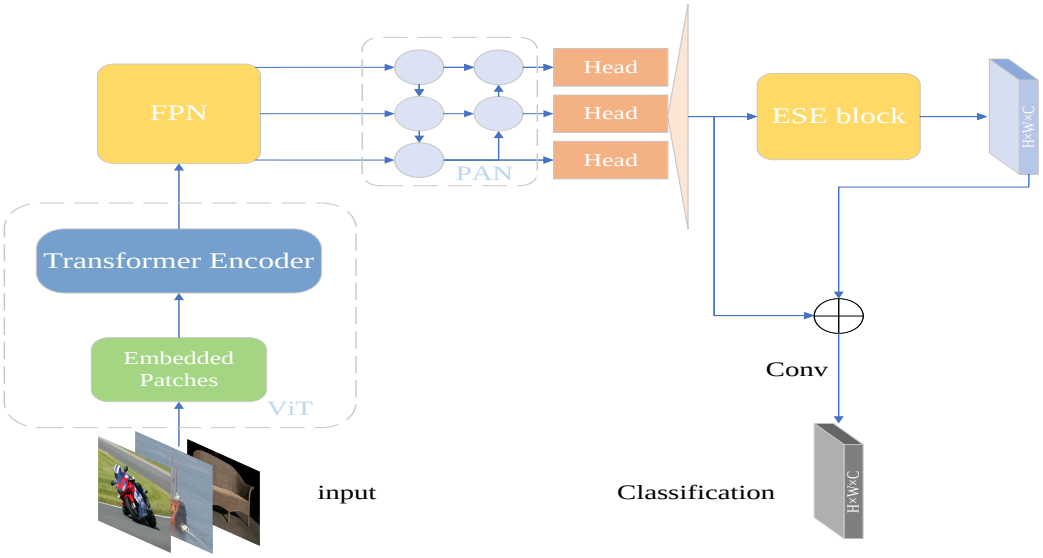
[7] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph, "Simple copy-paste is a strong data augmentation method for instance segmentation," in

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 2918–2928.

[8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[9] Wenyu Lv, Shangliang Xu, Yian Zhao, Guanzhong Wang, Jinman Wei, Cheng Cui, Yuning Du, Qingqing Dang, and Yi Liu, "Detrs beat yolos on real-time object detection," *arXiv preprint arXiv:2304.08069*, 2023.

- Representative image / diagram of the method(s):



3 Global Method Description

[* Indicates method used in competition test results.]

- Total method complexity:
  The training of this model is divided into three stages: self-monitoring

training, supervised training of object detection and semi-supervised training of object detection.

In the prediction, the multi-scale WBF strategy is adopted.

- Model Parameters:
  Supervised stage:
  A total of 160epoch were trained. AdamW optimizer was adopted, and the learning rate strategy was CosineDecay. Before that, 3 epochs were used to do Warmup, and the learning rate was set to 1e-4. The data enhancement used in training are Mosaic[6], copy paste[7], random distance, random expand, random crop, random flip and resize. Batch_size is 4*4.

- Run Time:
  In the supervised training stage, the total amount is 160epoch, which takes 23 hours.

  It takes 27 hours to train 128epoch in the semi-supervised stage.

  Under the single scale of the prediction stage, the test set length of the second stage is 7 min 24 s.

- Which pre-trained or external methods / models have been used:
  This scheme adopts the self-monitoring model CAE which is officially provided by paddle and trained on ImageNet.

- Training description :
  The first step is to select an appropriate object detection network. The networks compared here include FasterRCNN[8], RT-DETR[9] and PPYoloE. Under the backbone network of ResNet50, its performance is shown in Table 1 without any pre-training.

Table 1: Without any pre-training

Model	backbone	pretrained	epoch	mAP
FasterRCNN	ResNet50	×	12	12
RT-DETR	ResNet50	×	36	0.24
PPYoloE	ResNet50	×	72	0.23

The second step is to select the appropriate backbone network. In the backbone network stage, we mainly compared ResNet50 and CSP-ResNet50 with VIT. Because CSP-ResNet50 was completely trained by itself in the self-monitoring stage, the training time was too long and never ended, so it did not participate in the final selection. See Table 2 for details.

Table 2: Results after joining and training

Model	backbone	pretrained	epoch	mAP
PPYoloE	ResNet50	×	72	0.24
PPYoloE	ResNet50	√	72	0.264
PPYoloE	ViT-Base	×	72	0.262
PPYoloE	ViT-Base	√	72	0.285

Step 3, data-enhanced ablation experiment. At this stage, Pose, context, occlusion and weather with poor performance are added with different data enhancement strategies respectively. For pose, vertical flip is added on the basis of the original horizontal flip; Mosaic and copy-paste are added for context and occlusion; Badweather is added to weather, which includes five climate strategies: cloud, rain, snow, snow scene and fog. However, due to the time problem, it was not added to the final training. I believe that the model performance will have a better score after joining. See Table 3 for details.

The fourth step, semi-supervised training. We combine the training sets of official multiple tracks and the test sets of stage one into semi-supervised training data for the model to complete good semi-supervised training. As shown in Table 4.

- Testing description:
  In the testing stage, we completed the fusion strategy through WBF on the scale of 448 and 512. In order to ensure that the performance of some OOD features will not be degraded during semi-supervision, the model before semi-supervision is added in the final fusion. At the same time, according to their scores, the two models finally joined the configuration with a weight of 2: 5, and the optimal performance was obtained.The final results are shown in Table 5. Finally, we got a score of 0.494.

Table 3: The model is PPYoloE, the backbone is ViT-Base, and the results of different trick

Trick	epoch	IID-mAP	OOD-shape-mAP	OOD-pose-mAP	OOD-context-mAP	OOD-texture-mAP	OOD-occlusion-mAP	OOD-weather-mAP	mAP
Baseline	36	0.442	0.398	0.202	0.219	0.369	0.161	0.211	0.260
Flip	36	0.473	0.439	0.283	0.251	0.388	0.230	0.208	0.299
Mosaic +copy -paste	36	0.493	0.460	0.310	0.283	0.412	0.262	0.281	0.335
Bad-weather	36	0.450	0.441	0.267	0.255	0.368	0.219	0.225	0.296
Mosaic +Flip +copy -paste	108	0.541	0.499	0.380	0.334	0.450	0.337	0.329	0.388

Table 4: The second results of different trick when the model is PPYoloE and the backbone is ViT-Base

Trick	epoch	IID-mAP	OOD-shape-mAP	OOD-pose-mAP	OOD-context-mAP	OOD-texture-mAP	OOD-occlusion-mAP	OOD-weather-mAP	mAP
Mosaic +Flip +copy -paste	108	0.541	0.499	0.38	0.334	0.450	0.337	0.329	0.338
Dense-Teacher	128	0.565	0.545	0.459	0.411	0.483	0.597	0.387	0.480

- Quantitative and qualitative advantages of the proposed solution :
  The experiment is fast in training duration, and can achieve real-time monitoring performance without WBF fusion strategy, and keep a good mAP.

- Novelty of the solution and if it has been previously published:
  In this scheme, different data enhancement strategies are adopted according to different OOD conditions to ensure the stability of the model, and the semi-supervised method has better adaptability to OOD data and better portability, which can be used in different models.

Table 5: Other results of different trick when the model is PPYoloE and the backbone is ViT-Base

Trick	epoch	IID-mAP	OOD-shape-mAP	OOD-pose-mAP	OOD-context-mAP	OOD-texture-mAP	OOD-occlusion-mAP	OOD-weather-mAP	mAP
448 +512 +WBF	128	0.575	0.557	0.471	0.411	0.497	0.575	0.409	0.487
semi-supervised +supervised	128	0.574	0.560	0.474	0.417	0.498	0.588	0.412	0.492
Different-model-weights	128	0.578	0.561	0.477	0.419	0.501	0.591	0.417	0.494

4 Ensembles and fusion strategies

- Describe in detail the use of ensembles and/or fusion strategies (if any).: In the end, the experiment adopted the fusion strategy of multi-scale WBF.
- What was the benefit over the single method? : Compared with the detection results under a single scale, multi-scale can be more compatible with large and small objects, and WBF makes the determination of detection frames smoother and more reasonable.
- What were the baseline and the fused methods? : Baseline model is object detection model PPYoloE, and fusion strategy is multi-scale WBF.

5 Technical details

- Language and implementation details (including platform, memory, parallelization requirements) : This model uses python as the programming language and PaddlePaddle 2.4.0 as the development framework. It is best to run on the GPU of V100.
- Human effort required for implementation, training and validation?: The experiment was trained on the four V100 GPUs and verified on the single V100 GPU.

- Training/testing time? Runtime at test per image :
  In the supervised training stage, the total amount is 160epoch, which takes 23 hours. It takes 27 hours to train 128epoch in the semi-supervised stage. In the test, when WBF is not used, the reasoning time of a single picture is 0.03 seconds. The prediction time of a single image at multi-scale is 0.24 seconds. Because WBF is offline fusion, it is impossible to measure the duration.

- Comment the efficiency of the proposed solution(s)? :
  The scheme is a one-stage model in both the training stage and the forecasting stage, so it should be faster in efficiency.

6 Other details

- General comments and impressions of the OOD-CV challenge. :
  Out-of-Distribution(OOD) detection plays an important role in the stability and security of machine learning. The OOD task is to detect whether a machine learning model can correctly identify and classify all categories in the data set, especially those that do not belong to the training set. This kind of detection is very important to ensure the reliability and robustness of machine learning model in practical application. Therefore, we think that this challenge event is very necessary.

- Other comments:
  I hope the competition can get better and better. Great breakthroughs can also be made in this direction.