ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747353

# NOVEL INSTANCE MINING WITH PSEUDO-MARGIN EVALUATION FOR FEW-SHOT OBJECT DETECTION

*Weijie Liu*[1], *Chong Wang*[1,2*], *Shenghao Yu*[1], *Chenchen Tao*[1], *Jun Wang*[2] *and Jiafei Wu*[3]

[1] Faculty of Electrical Engineering and Computer Science, Ningbo University, China
[2] School of Information and Control Engineering, China University of Mining and Technology, China
[3] SenseTime Research, China

## ABSTRACT

Few-shot object detection (FSOD) enables the detector to recognize novel objects only using limited training samples, which could greatly alleviate model's dependency on data. Most existing methods include two training stages, namely base training and fine-tuning. However, the unlabeled novel instances in the base set were untouched in previous works, which can be re-used to enhance the FSOD performance. Thus, a new instance mining model is proposed in this paper to excavate the novel samples from the base set. The detector is thus fine-tuned again by these additional free novel instances. Meanwhile, a novel pseudo-margin evaluation algorithm is designed to address the quality problem of pseudo-labels brought by those new novel instances. The experimental results on MS-COCO dataset show the effectiveness of the proposed model, which does not require any additional training samples or parameters. Our code is available at: *https://github.com/liuweijie19980216/NimPme.*

***Index Terms***— Few-shot object detection, pseudo-labels, pseudo-margin evaluation

## 1. INTRODUCTION

Deep Convolutional Neutral Networks have achieved the great success in the extensive computer vision tasks, such as image classification [1, 2], object detection [3-10] and so on. Most of those algorithms need abundant training samples with fully annotated, which is time consuming and expensive. In order to address this issue, few-shot learning (FSL) has been proposed to enable model to learn new knowledge only using very few novel training samples. Many popular FSL methods are based on meta-learning [11-16], which establish the episode including support and query set to simulate the situation of insufficient training samples. These methods allow the model to be easily transferred to novel classes with a small number of samples.

However, the development of FSL mainly focus on image classification. Despite the factor that the box annotations are more difficult to obtain in the field of object detection, few-shot object detection (FSOD) gets relatively less attention. In recent years, most FSOD algorithms are based on the meta-learning [17-22] and transfer learning [23-25], which adapt two-stage training strategy. As shown in Fig. 1, the base detector is trained by plenty labeled training samples of base classes. Then, much less labeled training samples of balanced novel and base classes are used to fine-tune the novel detector. Generally, the number of available novel samples in the fine-tuning stage is positively related to the detection accuracy, e.g. 10-shot results are better 5-shot ones. However, the two-stage training strategy has ignored those unlabeled novel instances hidden in the base set, which can be re-used to enhance the detection performance. Such situation is quite common in real world that we have seen something many times before we know what it is.

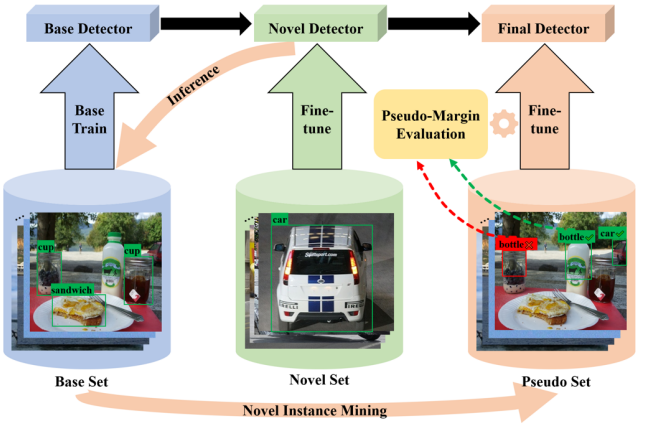


**Fig. 1.** Illustration of novel instance mining. The orange lines represent our method, using novel detector mining novel instances without true labels to fine-tune the final detector.

To address this issue, a new model of mining and re-using those novel instances is introduced in this paper. A similar idea has been proposed in the zero-shot object detection [26]. However, their novel instances were mined from the test set, which is improper to touch the test samples in the training stage. In contrast, a more reasonable way is applied in this paper to excavate the novel instances only from what is available during the training process, i.e. the base set. As shown by the orange lines of Fig. 1, the trained novel detector is used to infer the pseudo-labels (categories and bounding boxes) of novel classes in the base set. Then a novel set of training samples including novel instances with pseudo-labels is utilized to fine-tune the detector again.

Similar concept can be found in the field of semi-supervised object detection (SSOD) [27-29]. Some work [27]

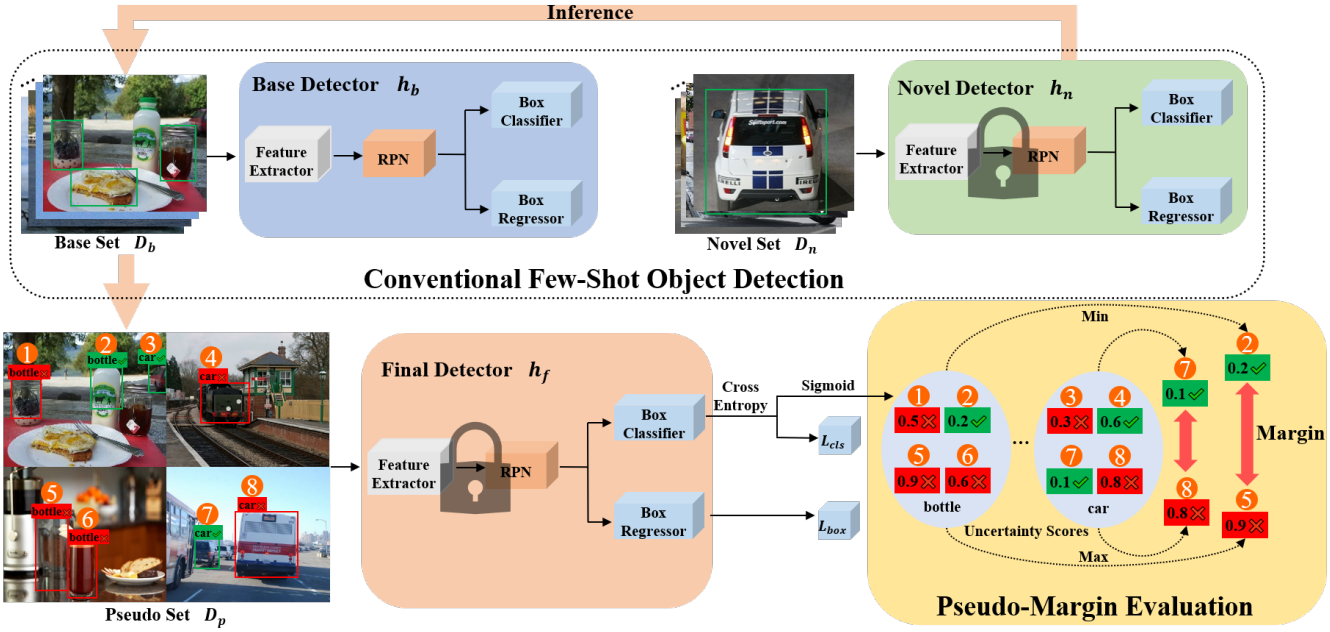


**Fig. 2.** Overview of the proposed model. The top left, top right and bottom parts indicates three training stages, namely base training, novel set fine-tuning and pseudo set fine-tuning with pseudo-margin evaluation. The lock icon means the parameters of this part are not updated.

has designed a teacher-student mutual learning mechanism to generate pseudo-labels on unlabeled dataset. The context information is used in [28] to filter noise pseudo-labels, while the quality of pseudo-labels is evaluated in [29] according to the image-level and regional-level, to be used in different stages. However, they all require a huge number of the unlabeled data, which is hard to be satisfied in FSOD.

Unlike most previous works that focus on generating more accurate pseudo-labels, a novel pseudo-margin evaluation (PME) model is proposed to introduce a new way to exploit the error-prone pseudo-labels by evaluating the uncertainty scores of both correct and incorrect pseudo-labels. It is capable to utilize both the true and false samples simultaneously to optimize the detector in the right direction, which is effective to avoid error reinforcement introduced by the incorrect pseudo-labels.

## 2. METHODOLOGY

Assume the base detector $h_b$ is first trained using base samples from the base set $D_b$. Then, the novel detector $h_n$ is fine-tuned using mixed novel and base samples from the novel set $D_n$. An extra fine-tuning stage is designed in this section to utilize excavated additional novel instances from $D_b$. Meanwhile, a pseudo-margin evaluation (PME) model is proposed to alleviate the influence of incorrect labels by exploiting the uncertainty in those new novel instances.

### 2.1. Overview of the proposed model

The workflow of the proposed model is shown in Fig. 2, which includes three training stages, namely base training, novel set fine-tuning and pseudo set fine-tuning. The first two training stages are the same as the previous FSOD methods. To be specific, the paraments $\theta_b = \{\theta_b^{(f)}, \theta_b^{(p)}\}$ and $\theta_n = \{\theta_n^{(f)}, \theta_n^{(p)}\}$ of the base detector $h_b(;\theta_b)$ and novel detector $h_n(;\theta_n)$ are respectively updated as,

$$\theta_b \leftarrow \theta_b + \gamma \frac{\partial(L_{rpn}+L_{cls}+L_{box})}{\theta_b}, \quad (1)$$

$$\theta_n^{(p)} \leftarrow \theta_n^{(p)} + \gamma \frac{\partial(L_{cls}+L_{box})}{\theta_n^{(p)}}, \quad (2)$$

$$\theta_n^{(f)} = \theta_b^{(f)}, \quad (3)$$

where $\gamma$ is the learning rate, $L_{rpn}$, $L_{cls}$ and $L_{box}$ represent the loss of Region Proposal Network (RPN), box classifier and box regressor. $\theta_n^{(f)}$ and $\theta_b^{(f)}$ denotes the parameters of the feature extractor, while $\theta_b^{(p)}$ and $\theta_n^{(p)}$ are the parameters of the box predictor including the classifier and regressor.

In the pseudo set fine-tuning stage, the base set $D_b$ is re-used to generate a new pseudo set $D_p$ containing plausible novel instances, which are inferred by $h_n(;\theta_n)$ from the 2nd stage. Such $D_p$ is then utilized to fine-tune the network again to obtain the final detector $h_f(;\theta_f)$. The proposed PME process is also conducted in this stage.

### 2.2. Novel instance mining

The key of our proposed model is to excavate more novel instances from the base set $D_b = \{(x_i, y_i^{(b)})\}_{i=1}^{N}$ for FSOD, where $x_i$ denotes the $i$-th image and $y_i^{(b)} = \{(c_{i,k}^{(b)}, l_{i,k}^{(b)})\}_{k=1}^{K_i}$ is its labels of $K_i$ base object boxes (the categories $c_{i,k}^{(b)}$ and locations $l_{i,k}^{(b)}$). Thus, the most straightforward way is to apply the novel detector $h_n(x_i;\theta_n)$ on $D_b$, which will get,

$$h_n(x_i;\theta_n) = \left\{\left(\hat{c}_{i,k}^{(n)}, \hat{l}_{i,k}^{(n)}\right)\right\}_{k=1}^{K_i}. \quad (4)$$

where $\hat{c}_{i,k}^{(n)}$ and $\hat{l}_{i,k}^{(n)}$ is the predicted novel category and its

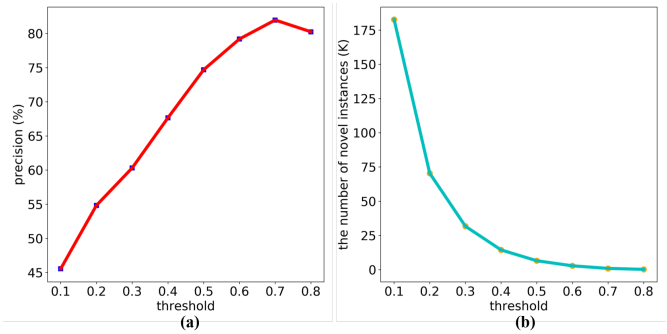


**Fig. 3.** The quality and quantity of excavated novel instances at different confidence thresholds.

location of *k*-th box in *i*-th image. Meanwhile a threshold $\delta$ is set to discard those boxes with low confidence $\hat{s}_{i,k}^{(n)}$ which is defined as,

$$\hat{s}_{i,k}^{(n)} = max\left(\hat{p}_{i,k}^{(n)}\right), \hat{p}_{i,k}^{(n)} \in \mathbb{R}^{C_n}, \tag{5}$$

$$\hat{p}_{i,k} = \{\hat{p}_{i,k}^{(b)}, \hat{p}_{i,k}^{(n)}, \hat{p}_{i,k}^{(bg)}\}, \tag{6}$$

where $C_n$ is the number of all novel classes, $\hat{p}_{i,k}^{(b)}$, $\hat{p}_{i,k}^{(n)}$ and $\hat{p}_{i,k}^{(bg)}$ dedicate probability distribution of base classes, novel classes and background given by the box classifier. Then the pseudo-label $\tilde{y}_i^{(n)}$ of novel instances for $x_i$ can be defined as,

$$\tilde{y}_i^{(n)} = \left\{\left(\hat{c}_{i,k}^{(n)}, \hat{l}_{i,k}^{(n)}\right) | \hat{s}_{i,k}^{(n)} > \delta\right\}_{k=1}^{K_i}. \tag{7}$$

Going through the whole dataset of $D_b$ to collect $N_p$ images containing pseudo-labels $\tilde{y}_i^{(n)}$, a new pseudo set $D_p$ can then be formed as,

$$D_p = \left\{\left(x_i, \tilde{y}_i^{(n)} \cup y_i^{(b)}\right)\right\}_{i=1}^{N_p}. \tag{8}$$

To avoid class imbalance, original annotations $y_i^{(b)}$ of base classes are retained in $D_p$. It is also worth noting that the pseudo-label $\tilde{y}_i^{(n)}$ is not always true. Its quality is highly dependent on the threshold $\delta$ and the accuracy of $h_n(x_i; \theta_n)$. As shown Fig. 3, a too small $\delta$ will result numerous misclassified boxes, while a too large one can rarely find useful novel objects. Instead of selecting an empirical value of $\delta$, a cleverer way is introduced in the next section.

### 2.3. Pseudo-margin evaluation

The obtained pseudo set $D_p$ can be directly used to fine-tune the final detector as the previous works [26]. However, the pseudo-labels $\tilde{y}_i^{(n)}$ are error prone, which makes the final detection performance be sensitive to the threshold $\delta$. With the prior knowledge that it is inevitable to have both correct and incorrect labels in $\tilde{y}_i^{(n)}$, it is feasible to utilize such difference between labels to fine-tune the final detector.

To estimate the quality of pseudo-labels, the uncertainty score $a_{i,j}^{(v)}$ of *j*-th novel Region of Interest (RoI) in the *i*-th image for the novel class $v$ can be evaluated by an activation after the cross-entropy,

$$a_{i,j}^{(v)} = Sigmoid(-\tilde{c}_{i,j}^{(v)} \log \hat{p}_{i,j}), \tag{9}$$

where $\tilde{c}_{i,j}^{(v)}$ dedicates pseudo-label of category, which is determined by choosing maximum Intersection over Union (IoU) of the *j*-th RoI with pseudo-label $\tilde{y}_i^{(n)}$ in the *i*-th image. And $\hat{p}_{i,j}$ dedicates the predicted class probability distribution of *j*-th RoI. Noting that the smaller $a_j$ is, the more likely its pseudo-label is correct.

Given a minibatch during the 3rd stage, for the involved $N_v$ RoIs of novel class $v$, the corresponded uncertainty score vector is defined as $A_v = \{a_1^{(v)}, \ldots, a_{N_v}^{(v)}\}$. As shown in Fig. 2, the maximum and minimum of uncertainty scores are chosen in pairs to perform pseudo-margin evaluation. Inspired by [30, 31], a new PME loss is defined in a similar form as the hinge loss,

$$L_{pme} = \sum_v max(0, \mu - max(A_v) + min(A_v)), \tag{10}$$

where $\mu$ is the margin that is set as 0.8 in our experiments.

It is worth noting that the maximum and minimum values of $A_v$ for each novel class are the best guess of the incorrect and correct pseudo-labels, respectively. Furthermore, the margin between them is used to guide the training of the box classifier, since a larger margin between true and false labels usually means better classification performance on the novel classes. Comparing to the directly use of all pseudo-labels, it is obviously more accurate and robust to the quality of $\tilde{y}_i^{(n)}$.

### 2.4. Pseudo set fine-tuning

With the proposed PME loss, the learnable weights $\theta_f^{(p)}$ of prediction head is fine-tuned using $D_p$ as,

$$\theta_f^{(p)} \leftarrow \theta_f^{(p)} + \gamma \frac{\partial(L_{cls} + L_{box} + L_{pme})}{\theta_f^{(p)}}. \tag{11}$$

Noting that $D_p$ contains both true labels of base classes and pseudo-labels of novel classes. Hence, the classification loss $L_{cls}$ is redefined as a weighted sum of the losses of base classes $L_{cls}^{(b)}$ and novel classes $L_{cls}^{(n)}$,

$$L_{cls} = \alpha L_{cls}^{(b)} + \beta L_{cls}^{(n)}, \tag{12}$$

where the weight $\alpha$ and $\beta$ are set as 0.9 and 0.1 in our experiments, respectively. Similarly, the regression loss $L_{box}$ only consider the RoIs from base classes, which is defined as smooth L1 loss.

## 3. EXPERIMENTS

In this section, the performance of the proposed model, namely novel instance mining with PME (N-PME), are evaluated on MS-COCO [32] dataset, which are also compared with the other state-of-the-art (SOTA) methods. Ablation studies and detailed analysis are also included.

### 3.1. Benchmarks and Setups

MS-COCO is a complex and realistic dataset including 80 categories, which is suitable to demonstrate the effectiveness of our model. Among them, 20 categories are belonging to

**Table 1**: the mAP of novel classes on MS-COCO (%).

| Shot | Method | AP | AP50 | AP75 |
|---|---|---|---|---|
| 10 | FSRW[17] | 5.6 | 12.3 | 4.6 |
| | MetaDet [18] | 7.1 | 14.6 | 6.1 |
| | Meta R-CNN[19] | 8.7 | 19.1 | 6.6 |
| | FRCN+ft-full [23] | 9.2 | 17.0 | 9.2 |
| | Retentive R-CNN [25] | 10.5 | - | - |
| | TFA w/fc [23] | 10.0 | 19.2 | 9.2 |
| | TFA w/cos [23] | 10.0 | 19.1 | 9.3 |
| | N-PME (Ours) | **10.6** | **21.1** | **9.4** |
| 30 | FSRW[17] | 9.1 | 19.0 | 7.6 |
| | MetaDet [18] | 11.3 | 21.7 | 8.1 |
| | Meta R-CNN[19] | 12.4 | 25.3 | 10.8 |
| | FRCN+ft-full [23] | 12.5 | 23.0 | 12.0 |
| | Retentive R-CNN [25] | 13.8 | - | - |
| | TFA w/fc [23] | 13.4 | 24.7 | 13.2 |
| | TFA w/cos [23] | 13.7 | 24.9 | 13.4 |
| | N-PME (Ours) | **14.1** | **26.5** | **13.6** |

the PASCAL VOC [33] dataset, which are regarded as novel classes in the experiments. Thus, the other 60 categories are chosen as base classes. Mean Average Precision (AP) with different Intersection over Union (IoU) with ground truth is used as performance evaluation criteria.

In the phase of base training, all training samples with base classes are used without any annotations of novel classes. For novel set fine-tuning, a given number of samples of base and novel classes are applied. For example, 10 training samples for every category are available in the setting for 10-shot. It follows the protocol in TFA [23]. In the pseudo set fine-tuning stage, those excavated samples in pseudo set $D_p$ are utilized, while the confidence threshold $\delta$ is set as 0.6.

### 3.2. Analysis of experimental results

In the experiments, the proposed N-PME is implemented on the widely used baseline TFA [23] in FSOD. As shown in Table 1, our N-PME achieves constantly higher accuracy than the baseline in term of AP, AP50 and AP75 in both of 10 and 30-shot settings. In the 10-shot setting, the SOTA method Retentive R-CNN [25] achieve 0.5% improvement by adding a sub-network on TFA. In contrast, our model can provide higher AP boost (0.6%) without any extra cost at inference stage. Moreover, the proposed N-PME have better performance (14.1%) than Retentive R-CNN (13.8%) in 30-shot setting as well. It is worth noting that our N-PME is a plug-and-play module, whose performance is limited by the baseline model (TFA [23]). Thus, the reported AP here is lower than SOTA methods based on meta-learning [21, 22].

There is an interesting phenomenon occurred in 10 and 30-shot setting. Compared to the significant improvement (2.0% and 1.6%) of AP50, the improvement of AP75 is relatively small (0.1% and 0.2%). The best guess is that it is limited by the accuracy of bounding box regression, since the novel instances are not used in the regression loss. This can also be observed in the visualization of detection results shown in Fig. 4. It can be seen that our model could detect more novel instances than TFA, but many box annotations are not accurate.

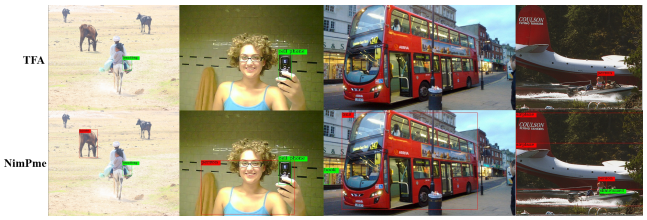


**Fig. 4.** The visualization comparison between TFA [23] and the proposed N-PME. (Red: novel classes, Green: base classes.)

**Table 2**: Ablation study of 10-shot setting on MS-COCO (%).

| Pseudo Set | PME | Novel Set | AP | AP50 | AP75 |
|---|---|---|---|---|---|
| | | | 10.0 | 19.1 | 9.3 |
| √ | | | 10.4 | 20.4 | 9.2 |
| √ | √ | | **10.6** | **21.1** | **9.4** |
| √ | √ | √ | **10.6** | **21.1** | **9.4** |

### 3.3. Ablation study for pseudo-margin evaluation

In order to express the characteristics of the proposed PME more specifically, a set of ablation study is shown in Table. 2. If the pseudo set $D_p$ is directly used on TFA to fine-tune the detector, the accuracy (AP, AP50 and AP75) has noticeably less improvement than the proposed model with PME. It shows the effectiveness of PME in dealing with the quality issue of pseudo labels.

Like SSOD, the novel set $D_n$ and pseudo set $D_p$ can be used together to fine-tune the detector. However, as the result shown in Table 2, there is no difference from the one only using $D_p$. The best guess is that the detector has converged on the novel set with a small number of training samples. Thus, it could not improve the performance further.

## 4. CONCLUSION

A new model of novel instance mining with pseudo-margin evaluation (N-PME) have been proposed to utilize the novel detector to excavate the novel instances from the base set. Meanwhile, the proposed N-PME can utilize the correct and incorrect pseudo-labels simultaneously to guide the network training by their uncertainty. The experimental results have shown that our model could effectively embed into conventional algorithms and improve their performance without any additional training samples and parameters. The proposed N-PME will be further improved to utilize pseudo-labels for regressor in our future.

## 5. ACKNOWLEDGMENT

This work was supported by the Scientific Innovation 2030 Major Project for New Generation of AI, Ministry of Science and Technology of the People's Republic of China (No. 2020AAA0107300), Zhejiang Provincial Natural Science Foundation of China (No. LY20F030005) and National Natural Science Foundation of China (No. 61603202).

## 6. REFERENCES


[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *NIPS*, pp. 1097-1105, 2012.

[2] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, pp. 770-778, 2016.

[3] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *CVPR*, pp. 779-788, 2016,

[4] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," in *CVPR*, pp. 6517-6525, 2017.

[5] W. Liu, et al., "Ssd: Single shot multibox detector," in *ECCV*, pp. 21-37, 2016.

[6] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *CVPR*, pp. 580-587, 2014.

[7] R. Girshick, "Fast R-CNN," in *ICCV*, pp. 1440-1448, 2015.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *NIPS*, pp. 91-99, 2015.

[9] T. Lin, et al., "Feature Pyramid Networks for Object Detection," in *CVPR*, pp. 936-944, 2017.

[10] Q. Mao, C. Wang, S. Yu, Y. Zheng, and Y. Li, "Zero-Shot Object Detection With Attributes-Based Category Similarity," In *TCAS-II*, pp. 921-925, 2020.

[11] A. Santoro, et al., "Meta-learning with memory-augmented neural networks," in *ICML*, pp. 1842-1850, 2016.

[12] S. Ravi, and H. Larochelle, "Optimization as a Model for Few-Shot Learning," in *ICLR*, 2017.

[13] C. Finn, P. Abbeel, and S. Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," in *PMLR*, pp. 1126-1135, 2017.

[14] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *NIPS*, pp. 4080-4090, 2017.

[15] F. Sung, et al., "Learning to Compare: Relation Network for Few-Shot Learning," in *CVPR*, pp. 1199-1208, 2018.

[16] O. Vinyals, et al., "Matching Networks for One Shot Learning," in *NIPS*, pp. 3637-3645, 2016.

[17] B. Kang, et al., "Few-Shot Object Detection via Feature Reweighting," in *ICCV*, pp. 8419-8428, 2019.

[18] Y. Wang, D. Ramanan and M. Hebert, "Meta-Learning to Detect Rare Objects," in *ICCV*, pp. 9924-9933, 2019.

[19] X. Yan, et al., "Meta R-CNN: Towards General Solver for Instance-Level Low-Shot Learning," in *ICCV*, pp. 9576-9585, 2019.

[20] W. Liu, et al., "Dynamic Relevance Learning for Few-Shot Object Detection," arXiv preprint arXiv: 2108.02235, 2021.

[21] H. Hu, S. Bai, A. Li, J. Cui and L. Wang, "Dense Relation Distillation with Context-aware Aggregation for Few-Shot Object Detection," in *CVPR*, pp. 10180-10189, 2021.

[22] G. Zhang, Z. Luo, K. Cui, S. Lu, "Meta-DETR: Image-Level Few-Shot Object Detection with Inter-Class Correlation Exploitation," arXiv preprint arXiv:2103.11731, 2021.

[23] X. Wang, et al., "Frustratingly Simple Few-Shot Object Detection," in *ICML*, pp. 9919-9928, 2020.

[24] J. Wu, S. Liu, D. Huang, Y. Wang, "Multi-scale Positive Sample Refinement for Few-Shot Object Detection," in *ECCV*, pp. 456-472, 2020.

[25] Z. Fan, Y. Ma, Z. Li and J. Sun, "Generalized Few-Shot Object Detection Without Forgetting," in *CVPR*, pp. 4527-4536, 2021.

[26] S. Rahman, S. Khan and N. Barnes, "Transductive Learning for Zero-Shot Object Detection," in *ICCV*, pp. 6081-6090, 2019.

[27] Y-C, Liu, et al., "Unbiased teacher for semi-supervised object detection," in *ICLR*, 2021.

[28] P. Tang, C. Ramaiah, R. Xu, and C. Xiong, "Proposal learning for semi-supervised object detection," arXiv preprint arXiv:2001.05086, 2020.

[29] Z. Wang, et al., "Data-Uncertainty Guided Multi-Phase Learning for Semi-Supervised Object Detection," in *CVPR*, pp. 4568-4577, 2021.

[30] W. Sultani, C. Chen and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," in *CVPR*, pp. 6479-6488, 2018.

[31] S. Yu, C. Wang, Q. Mao, Y. Li and J. Wu, "Cross-Epoch Learning for Weakly Supervised Anomaly Detection in Surveillance Videos," in *IEEE Signal Process. Lett*, pp. 1-5, 2021.

[32] T. -Y. Lin et al., "Microsoft COCO: Common Objects in Context," in *ECCV*, pp. 740-755, 2014.

[33] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," in *IJCV*, pp. 303-338, 2010.