A Semi-Supervised Learning Method for Masks Detection

Written by XMU Students[1]
Kaixuan Wu*31520211154092, Yang Xu*31520211154099, Zhaorun Wu*31520211154094, Ying Wang*31520211154089, Pengbo Yan*31520211154101

[1]Electronic Information Major, Department of Artificial Intelligence, School of Information, Xiamen University
[1]Electronic Information Major, Artificial Intelligence Research Institute, Xiamen University

Abstract

Since the outbreak of the COVID-19 epidemic, wearing masks has become an important protective measure for people going out. Many workplaces require employees to always wear masks to work. Therefore, the use of computer vision technology to detect whether people wear masks in real time has become one of the hotspots in target detection research. The current object detection framework based on supervised learning demand plenty of laborious manual annotations, which may be impractical in practical applications. Semi-supervised target detection (SSOD) can effectively use unlabeled data to improve model performance, which is of great significance for the application of object detection models. In this paper, we propose an effective SSOD framework, which builds a object detection model based on yolov5, uses pseudo-labels and weak-strong data enhancement to build the consistency loss of unlabeled data, and uses the Mean Teacher to train the teacher-student model to reduce the influence of false label noise. We tested it on the MaskDataset we built, and compared it with the supervised method when the labeled data was insufficient. Our model has a significant improvement: 41.5 mAP at 2% protocol, 8.8 mAP at5% protocol, 3.0 mAP for 10% protocol, and 2.3 mAP at 20% protocol, upon the supervised baselines, as shown in Figure 1.

Introduction

In early 2020, as COVID-19 spread across the globe, the issue of prevention became a common concern of citizens, and the use and popularity of face masks once again reached an unprecedented peak. So far, localized outbreaks of the epidemic remain inevitable. With the epidemic, wearing a mask in public is not only a moral obligation, but also becomes a legal one. In recent years, deep learning, with its successful application in speech recognition and computer vision, has made it a new direction in machine learning. The mask detection system in public places can automatically detect the wearing of masks through the existing monitoring equipment and the related method of machine vision, which can realize the rapid detection of the wearing of masks of people in public places, and carry out intelligent supervision through real-time monitoring and automatic alarm measures to alleviate the shortage of manpower in streets and communities at the grassroots level and improve the efficiency of epidemic prevention and control and the information level of supervision.

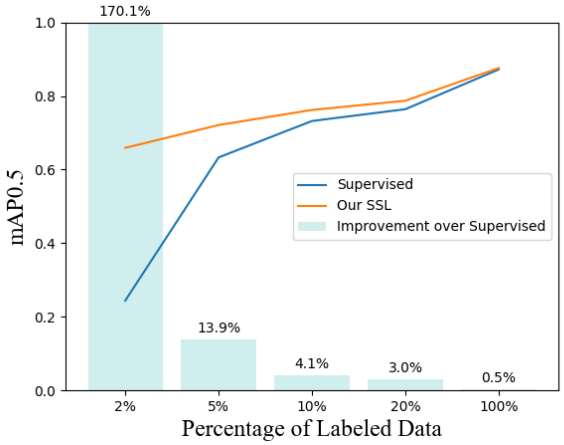

Figure 1: Our semi-supervised learning framework for object detection, consistently improves upon supervised baselines and those with data augmentation using different amount of labeled training data on MaskDataset.

Deep neural networks usually achieve their powerful performance through supervised learning. However, the model performance obtained by training depends heavily on annotated training data, mainly on the scene and scale of data. Using large amounts of labeled data can be costly, but we have very easy access to large amounts of unlabeled data, and we need to make full use of unlabeled data to improve the performance of the model. Applying the semi-supervised method to the field of object detection can greatly reduce the need for labeled datasets.

Mask detection refers to detecting whether a person is wearing a mask and whether he or she is wearing it in the correct position. There have been many studies on this specific problem, but the vast majority of them are based on supervised learning, and research based on semi-supervised

*These authors contributed equally.

methods is still limited. For this reason, our team proposes a semi-supervised mask-wearing detection.

Related work

Data augmentation is a strategy that is used to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from existing data. Data augmentation techniques such as cropping, padding, and horizontal flipping are commonly used to train large neural networks, especially gradually become a major impetus on semi-supervised learning. The complexity of data augmentations for object detection is much higher than image classification(Zoph et al. 2019), since the bounding box annotations. Many works have proposed data augmentation for object detection, such as MixUp(Zhang et al. 2017), CutMix(Yun et al. 2019). GAN(Goodfellow et al. 2014) based Augmentation which is used to translate images for data augmentation, such as CycleGan(Zhu et al. 2017), StarGan(Choi et al.).

Semi-supervised learning has made significant progress in image classification, which effectively utilizes a lot of unlabeled data. It's mainly divided into three main methods as following. Consistency regularization is widely used in the field of SSL. It refers to the idea that a model's response to an input should remain consistent, when perturbations are used on the input or the model. Pseudo labels are artificial labels generated by the model itself and are used to further train the model. Pseudo-label methods is to label unlabeled data which usually obtained by self-training, we can also use co-training to train multiple models at the same time and perform mutual verification to improve the quality of pseudo-labels. The third one is data augmentation and augmentation anchoring. Mean Teacher (Tarvainen and Valpola 2017) averages model weights instead of label predictions. It improves test accuracy and enables training with fewer labels than Temporal Ensembling.MixMatch (Berthelot et al. 2019b) works by generating pseudo-labels for the unlabeled data which obtained by averaging multiple sets of weak enhancements, mix them up with labeled data and train them with supervised techniques. Augmentation Anchoring is first proposed by ReMixMatch (Berthelot et al. 2019a) and further developed in FixMatch (Sohn et al. 2020). It is a form of consistency regularization that involves applying different levels of perturbations to the input. A model's response to a slightly perturbed input is regarded as the "anchor", and we try to align model's response to a severely perturbed input to the anchor. The current methods basically combine the above methods to use them together in order to achieve the best results.

Object detection is an important computer vision task used to detect instances of visual objects of certain classes (for example, humans, animals, cars, or buildings) in digital images such as photos or video frames. The state-of-the-art object detection methods can be categorized into two main types: One-stage vs. two-stage object detectors. In two-stage object detectors, the approximate object regions are proposed using deep features before these features are used for the classification as well as bounding box regression for the object candidate. For example, Faster R-CNN (Ren et al. 2015) proposes an RPN candidate frame generation algorithm based on Fast R-CNN, which greatly improves the speed of target detection. However, One-stage detectors predict bounding boxes over the images without the region proposal step. This process consumes less time and can therefore be used in real-time applications. YOLO (Redmon et al. 2016; Redmon and Farhadi 2017, 2018; Bochkovskiy, Wang, and Liao 2020) redefines object detection as a regression problem. It applies a single convolutional neural network (CNN) to the entire image, divides the image into grids, and predicts the class probability and bounding box of each grid. SSD (Liu et al. 2016) adds the anchor mechanism of Faster R-CNN on the basis of YOLO, which is equivalent to adding the mechanism of region suggestion on the basis of regression.

The semi-supervised object detection method mainly has two directions: Consistency based Learning and Pseudo-label based Learning. Recently, Qiang Zhou et al. improved STAC and proposed Instant-Teaching (Zhou et al. 2021). Instant-Teaching mainly proposed a co-rectify scheme to solve the problem of pseudo label confirmation bias (the cumulative effect of noise pseudo label errors). Zhenyu Wang et al. (Wang et al. 2021) proposed a multi-stage learning semi-supervised object detection learning algorithm, mainly to solve the problem of label noise overfitting caused by the strong fitting ability of deep network. Qize Yang et al. (Yang et al. 2021) proposed a semi-supervised target detection algorithm based on Mean Teacher. Interactive form of self-training,solving the problem that the previous method using pseudo label ignored the difference between the detection results of the same image in different iterations.Unbiased Teacher (Liu et al. 2021) jointly trains a student and a gradually progressing teacher in a mutually-beneficial manner. Together with a class-balance loss to downweight overly confident pseudo-labels, it consistently improves state-of-the-art methods by significant margins on some datasets.

Methodology

Our goal is to solve the problem of object detection in a semi-supervised setting and apply it to the specific scene of mask detection. We have a set of labeled images $D_s = \{x_i^s, y_i^s\}_{i=1}^{N_s}$ and s set of unlabeled images $D_u = \{x_i^u\}_{i=1}^{N_s}$ for training. N_s and N_u are the number of supervised and unsupervised data. For each labeled image x_s the annotations y_s contain locations, sizes, and object categories of all bounding boxes.

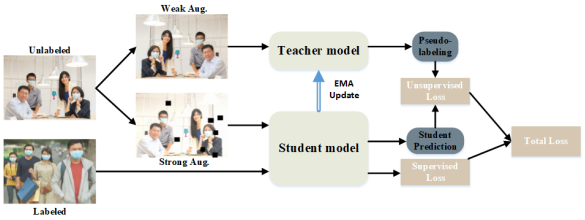

Figure 2: Semi-supervised mask detection framework

Overview As shown in Figure 2, our mask detection consists of two parts, One part uses labeled data for super-

vised training, and the other part uses unlabeled data for semi-supervised training. Before the Teacher-Student Mutual Learning stage, we use the pre-trained model of yolov5 to initialize the model with labeled data, then we duplicate the initialization model into two models (Teacher and Student models).

In each round of iterative process, the labeled images are sent to the student model, and the supervision loss L_{sup} is calculated based on the prediction result of the student model and the label. We perform two different degrees of enhancement on unlabeled images. The weakly enhanced images are sent to the teacher model, and pseudo-labels are generated based on the prediction results of the teacher model; the strongly enhanced images are sent to the student model, based on the student model's Calculate unsupervised loss L_{unsup} between prediction results and pseudo-labels. In the end we get the loss function of mask detection as follows:

$$L = L_{sup} + \alpha L_{unsup} \tag{1}$$

Mask detection based on yolov5 In order to accurately determine whether the person in the image wears a mask, it is first necessary to find the position of the face in the image, and then perform two-class recognition of the face in the bounding box. We refer to the yolov5 framework and set up loss function in three directions for mask detection: position prediction loss L_{pos}, confidence loss L_{con}, and classification loss L_{cls}. The L_{pos} is used to calculate the loss of bounding box in order to shorten the distance between the prediction box and the target box as soon as possible. As in yolov5, we use GIOU loss:

$$GIoU = IoU - \frac{|C \backslash (A \cup B)|}{|C|} \tag{2}$$

The above formula means that taking two bounding boxes A (true bounding box) and B (predicted bounding box), we find a smallest closed box C, so that C can include A and B, and then calculate the ratio of the area of C that does not cover A and B to the total area of C , And then subtract this ratio from the IoU of A and B. When GioU is used as the distance, the position prediction loss can be expressed as:

$$L_{pos} = 1 - GIoU \tag{3}$$

Both the confidence loss and the classification loss use the cross-entropy loss function, where the confidence loss is divided into the confidence loss of the box containing face and the confidence loss of the box not containing face:

$$L_{con} = \sum_{i=0}^{K\times K}\sum_{j=0}^{M} I_{ij}^{obj}[\hat{C}_i log(C_i) + (1-\hat{C}_i) log(1-C_i)] + \lambda_{noobj} \sum_{i=0}^{K\times K}\sum_{j=0}^{M} I_{ij}^{noobj}[\hat{C}_i log(C_i) + (1-\hat{C}_i) log(1-C_i)] \tag{4}$$

$$L_{cls} = \sum_{i=0}^{K\times K} I_{ij}^{obj} \sum_{c\in classes} [\hat{p_i(c)} log(p_i(c)) + (1-\hat{p_i(c)}) log(1-p_i(c)] \tag{5}$$

There are K×K grids in total, and each grid generates M anchors. Each anchor will get a corresponding bounding box through the network, and finally form K×K×M bounding boxes. If there is no face in the box, only the confidence loss of the box is calculated. The final supervision loss can be expressed as:

$$L_{sup} = L_{pos} - L_{con} - L_{cls} \tag{6}$$

Student Learning with Pseudo-Labeling To make full use of unlabeled data to train a model, a common approach is to generate pseudo-labels for unlabeled data. In the classification problem, it is considered a feasible and effective method to select data with high confidence pseudo-labels and continue to train the model. Similarly, we apply pseudo-labels to target detection and set a confidence threshold σ for predicting the bounding box, so as to select higher-quality prediction results as pseudo-labels to reduce the impact of noise caused by incorrect predictions. At the same time, in order to solve the problem of repeated box prediction, we eliminate repeated predictions by applying class-level non-maximum suppression (NMS) before using the confidence threshold.

In order to improve the quality of pseudo-labels, a common approach in semi-supervised classification tasks is to add consistency loss. For this reason, we perform weak enhancement and strong enhancement on unlabeled images respectively, and hope that the images after weak enhancement and strong enhancement will get consistent prediction results. For training, the image goes through random flipping and resizing as the weak augmentation. Upon the same weakly augmented image, we further randomly change the color, sharpness, contrast, add Gaussian noise and apply Cutout (DeVries and Taylor 2017). Using strongly enhanced images will increase the difficulty of the student's task and encourage it to learn better representations. In contrast, using weak enhancements for teachers can increase the chances of teachers generating correct pseudo-labels.

For the pseudo-labeled images, they are sent to the student model for training like the labeled images, and the unsupervised loss is calculated. The only difference is that the label of the image is replaced with the pseudo label generated by the teacher model:

$$L_{unsup} = L_{pos}^{u} - L_{con}^{u} - L_{cls}^{u} \tag{7}$$

Then the student model updates its own weight parameters based on backpropagation, and the learning rate is γ :

$$\theta_s \leftarrow \alpha\theta_s + \gamma \frac{\partial(L_{sup} + \alpha L_{unsup})}{\partial \theta_s} \tag{8}$$

Update teacher model In order to obtain higher-quality pseudo-labels and reduce the influence of noise data on the model during the training process, we use the EMA method to gradually update the teacher model. The teacher model obtained in this way can be regarded as the ensemble of the Student models in different training iterations.

$$\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s \tag{9}$$

θ_s represents the parameters of the student model, θ_t represents the parameters of the teacher model, α is a smoothing coefficient hyperparameter.

Experiments

Datasets

We test the efficacy of our proposed method on our own datasets, we call it MaskDataset, which contains more than 9k labeled images and 2 object categories(mask, unmask) for training and testing. 7,952 of them are from Baidu ai studio. Among the rest of 1314 images, the data of those having masks are partly crawled from Baidu through keyword search, and partly obtained by cutting frames from bilibili videos; the data of those not having masks are pieced together from the data sets of 'CASIA-FaceV5', 'WIDER-FACE', 'VOC2007' datasets.

Data Characteristics Through the analysis of the data, we found the following characteristics of the data: the distribution of data between categories is unbalanced, the number of 'mask' in the dataset is 5,464, the number of 'unmask' is 13,651, and the ratio of 'mask' to 'unmask' is 2:5. the aspect ratio of the bounding box is basically distributed between 0.7 and 2.0.

Data preprocessing The data preprocessing consists of three main parts: data format conversion, cleaning, and normalization operations. Cause the annotated data in the dataset have different formats, we need to unify the labeling format. We used PASCAL VOC (Everingham et al. 2010) format in experiment. In the data cleaning process, we checked the image data in the dataset and screened out the data with insignificant features and duplicates.

Dataset protocol We randomly sampled 8,000 images for the training set, and the rest of the data was used for testing and validation. In addition, we randomly sampled 2, 5, 10 and 20% of labeled training data as a labeled set and use the total of labeled training data as an unlabeled set. For these experiments, we create 5 data folds. 2% protocol contains 160 labeled images randomly selected from the training set. 5, 10, 20% protocol datasets are constructed in a similar way. We also experimented on the full data.

Evaluation Protocol The accuracy of detection in each category is very important. Therefore, in this paper, average precision (AP) and mean average precision (mAP) are selected as evaluation metrics for the object detection algorithm. We employed precision P and recall R as metrics, they are defined as follows:

$$P(class) = \frac{TP}{TP + FP} \tag{10}$$

$$R(class) = \frac{TP}{TP + FN} \tag{11}$$

Implementation Details

Our implementation is based on the Yolov5 which used CSPDarknet53 (Wang et al. 2020) backbone for object detector models. We use confidence threshold σ = 0.85 for pseudo labels and $\lambda_u = 0.5$ for unsupervised loss. For the data augmentation, we apply random horizontal flip for weak augmentation and randomly add color jittering, grayscale, Gaussian blur, and cutout patches for strong augmentations. The model is trained for 100 epochs, of which 50 were in Burn In stage. With SGD training, the learning rate is initialized to 0.01. With the one cycle learning rate scheduler, the learning rate at the last epoch is 0.002. The weight decay and the momentum are set to 0.0005 and 0.937, respectively. The batch size of supervised and unsupervised are both 16. We apply α = 0.999 as the EMA rate in Burn In stage, and α = 0.8 as the EMA rate for teacher model to update during the last 50 epochs. We use AP50 as evaluation metric, and the performance is evaluated on the Teacher model.

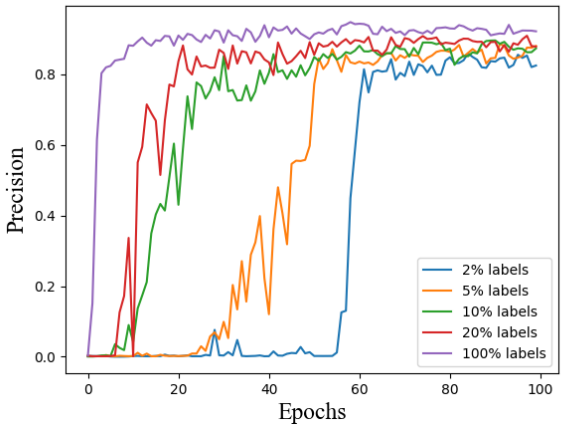

Figure 3: Precision on different dataset protocol at different training epochs.

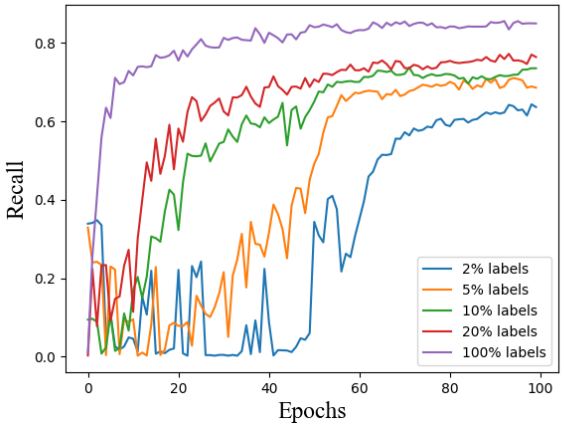

Figure 4: Recall on different dataset protocol at different training epochs.

Results

We ran the model on a NVIDIA GeForce RTX 2080, 2080 Ti and 3070, each experiment trained for 100 epochs. Since semi-supervised object detection has not been widely studied yet, we mainly compare our models with the supervised models (models trained with labeled data only) for various experimental protocols. We used the trained model to inference on the same test set. The mAP0.5 of the detection result

Methods	2%	5%	10%	20%	100%
Supervised	24.4%	63.3%	73.2%	76.4%	87.2%
Our SSOD	**65.9%**	**72.1%**	**76.2%**	**78.7%**	**87.6%**

Table 1: Train the same epochs, comparison in mAPs for different methods on MaskDataset. We report both mAPs at IoU=0.5(a standard metric), over 4 data folds for 2, 5, 10 and 20% protocols."Supervised" refers to models trained on labeled data only, which then are used to provide pseudo labels for Our SSOD. We train Our SSOD with data augmentation for unlabeled data.

of our algorithm is shown in Tabel 1. We confirm that data augmentation can indeed improve the performance of supervised learning. 41.5 mAP at 2% protocol, 8.8 mAP at 5% protocol, 3.0 mAP for 10% protocol, and 2.3 mAP at 20% protocol, upon the supervised baselines. The precision and recall during training are shown in Figure 3 and Figure 4. Experimenting on images using our trained semi-supervised model, some of the detection results are shown in Figure 5 to 8.

Conclusion

In this paper, we proposed a simple semi-supervised object detection method for mask wear detection. Our trained models improve the detector by leveraging a student model, and a teacher model which is continuously updated by the student model through the exponential moving average strategy. We demonstrated the effectiveness of our model design and achieves satisfactory performance. In this experiment, there is category imbalance in the data set used, so semi-supervised object detection in unbalanced scenarios is a worthy direction of study.

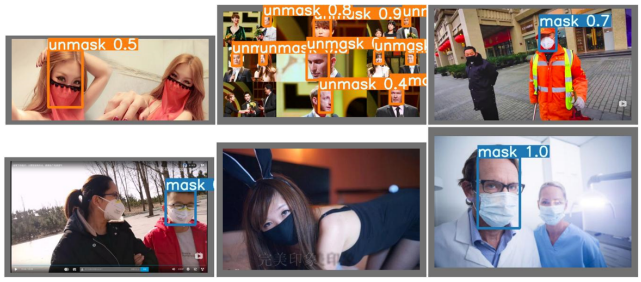

Figure 5: Results of 2% Labeled Data

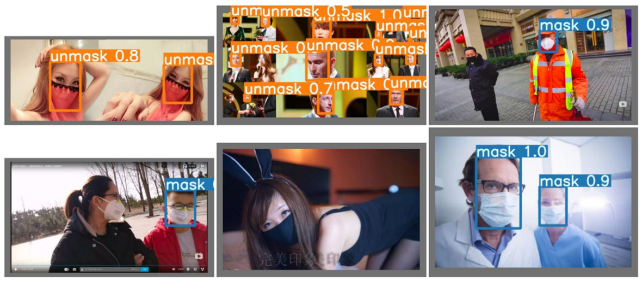

Figure 6: Results of 5% Labeled Data

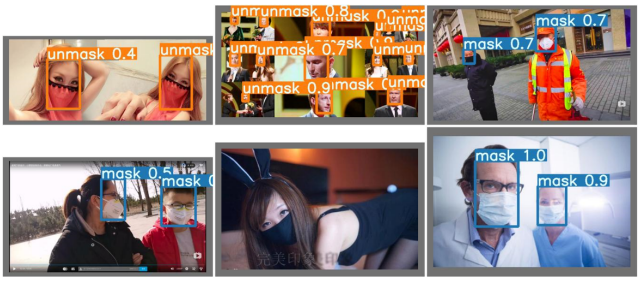

Figure 7: Results of 10% Labeled Data

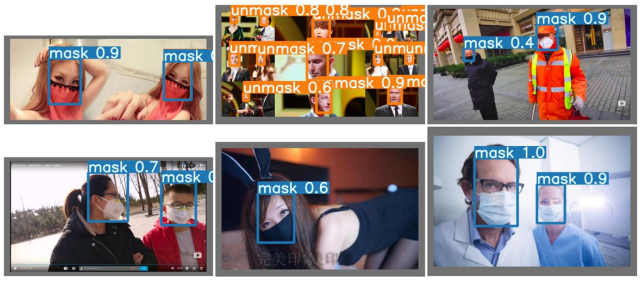

Figure 8: Results of 20% Labeled Data

References

Berthelot, D.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Sohn, K.; Zhang, H.; and Raffel, C. 2019a. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. 2019b. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*.

Bochkovskiy, A.; Wang, C.-Y.; and Liao, H.-Y. M. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

Choi, Y.; Choi, M.; Kim, M.; Ha, J. W.; and Choo, J. ???? StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation.

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37. Springer.

Liu, Y.-C.; Ma, C.-Y.; He, Z.; Kuo, C.-W.; Chen, K.; Zhang, P.; Wu, B.; Kira, Z.; and Vajda, P. 2021. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Redmon, J.; and Farhadi, A. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7263–7271.

Redmon, J.; and Farhadi, A. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28: 91–99.

Sohn, K.; Berthelot, D.; Li, C.-L.; Zhang, Z.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Zhang, H.; and Raffel, C. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*.

Wang, C.-Y.; Liao, H.-Y. M.; Wu, Y.-H.; Chen, P.-Y.; Hsieh, J.-W.; and Yeh, I.-H. 2020. CSPNet: A new backbone that can enhance learning capability of CNN. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 390–391.

Wang, Z.; Li, Y.; Guo, Y.; Fang, L.; and Wang, S. 2021. Data-Uncertainty Guided Multi-Phase Learning for Semi-Supervised Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4568–4577.

Yang, Q.; Wei, X.; Wang, B.; Hua, X.-S.; and Zhang, L. 2021. Interactive self-training with mean teachers for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5941–5950.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond Empirical Risk Minimization.

Zhou, Q.; Yu, C.; Wang, Z.; Qian, Q.; and Li, H. 2021. Instant-Teaching: An End-to-End Semi-Supervised Object Detection Framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4081–4090.

Zhu, J. Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *IEEE*.

Zoph, B.; Cubuk, E. D.; Ghiasi, G.; Lin, T. Y.; and Le, Q. V. 2019. Learning Data Augmentation Strategies for Object Detection.