

Brands Logo Recognition with Semi-Supervised Learning using a Robust YOLOv8 Detection Model

Ali Usman

ICT Convergence System
Engineering Department
Chonnam National University
Gwangju, South Korea
usman4293@gmail.com

Seungmin Oh

ICT Convergence System
Engineering Department
Chonnam National University
Gwangju, South Korea
osm5252kr@gmail.com

Junghoon Lee

ICT Convergence System
Engineering Department
Chonnam National University
Gwangju, South Korea
qlfxkdla33@gmail.com

Jinsul Kim

ICT Convergence System
Engineering Department
Chonnam National University
Gwangju, South Korea
jsworld@jnu.ac.kr

Abstract—Semi-supervised object recognition has emerged as a prominent area of research within computer vision. It offers the potential to significantly decrease the necessity for costly bounding-box annotations. Although there has been considerable success, the current advancements primarily concentrate on two-stage detection networks such as Faster RCNN, while study pertaining to single-stage detectors receives limited attention. This paper centers its attention on semi-supervised learning applied to the advanced and widely adopted single-stage detection network YOLOv8. Our method uses only a minimal amount of labeled data to successfully carry out the training, where we have explored various approaches like data augmentation, student-teacher network, pseudo labeling, and transfer learning. Furthermore, we have refined YOLOv8 implementation to optimize the advantages offered by semi-supervised learning. To validate our approach, we performed extensive experiments on the challenging OpenLogo Dataset, which contains 27,000 images across a total of 352 classes. The results were obtained with a limited amount of labeled data and a substantial amount of unlabeled data.

Keywords— *semi-supervised learning, YOLOv8, student teacher network, brands logo recognition.*

I. INTRODUCTION

The recent year has seen significant advancements in object detection, thanks to a multitude and innovative methods [1], [2], [3], [4]. While achieving remarkable success, the use of object detection has long been hindered by the requirement for annotations. To address this challenge, several research studies [5], [6], [7] have been dedicated to the field of semi-supervised object detection.

Taking inspiration from the progress made in image classification [8], [9], recent initiatives have also turned to the approach of teacher-student learning for semi-supervised object detection [5], [6], [7]. The core principle involves the utilization of a teacher network to produce pseudo labels for enhancing the performance of the student network, and to ensure the alignment of both networks robust and subtle data augmentations are separately applied, as discussed in [10], [11].

Nevertheless, these approaches primarily concentrate on two-stage detection networks such as Faster RCNN [2], with limited exploration in the widely adopted one-stage models like the YOLO models [12], [3], [4]. It can turn out to be an optimal choice to directly apply the current teacher-student method to

the one-stage object detector because of the enormous difference between one-stage and two-stage methods.

In this article, we introduced a teacher-student learning method designed specifically for single-stage object detection. Our base model YOLOv8 [15], which is one of the cutting-edge detection models, incorporates several training strategies within its framework, including Exponential Moving Average (EMA) [8], data augmentations, rate decay, and cosine-based learning. Simultaneously, we made adjustments to certain hyperparameters based on the semi-supervised training.

II. RELATED WORK

A. Semi-Supervised Object Detection

Object detection [1], [2], [3], [13] in the context of deep learning approaches can be roughly divided into two main categories: one-stage and two-stage. To be more precise, two-stage techniques [1], [2], [14] generate an acceptable amount of possible object parts first, subsequently, aggregate the features of these parts or regions to predict the appropriate bounding box and classes. The adaptability and performance of two-stage algorithms are frequently higher than those of single-stage approaches, however, their inference is a slower contrary to single-stage methods [3], [4] which predict the bounding box and classes of directly depending on the feature maps.

The method of pseudo-label based semi-supervised learning, as proposed in earlier work [16], involves utilizing the model's own predictions as definitive labels for guiding the semi-supervised training process, and in recent studies, researchers have turned to the approach of teacher-student learning [8], [9], [17]. Specifically, the Mean Teacher approach, as introduced in previous studies [8] implements data augmentation on the student network and compute the consistency loss and to enhance training stability and incorporates EMA to update the teacher's parameters based on the student's parameters. This strategy serves as an effective measure against confirmation bias, as discussed in [6].

Recent studies [5] stand out as a prominent example of a teacher-student based approach for semi-supervised object detection. It split the semi-supervised learning process into two stages: the initial pseudo-labeling phase for unlabeled data, followed by the subsequent re-training phase predicated on these pseudo-labels. In this research, our main objective is to

investigate the applicability of the teacher-student learning approach towards the robust single-stage detection model [15]. Additionally, we investigate the complex training methods frequently employed in single-stage models to evolve with the concept of teacher-student learning.

III. METHODOLOGY

The proposed architecture in Fig.1 is composed of two networks with similar settings, which are the student and teacher detection models. The teacher model plays a pivotal role in generating pseudo labels, which subsequently guide the training of the other model, alongside the ground truth. The optimization process for the student network is explicitly stated by.

$$\mathcal{L} = \mathcal{L}_{SUP} \lambda + \mathcal{L}_{UNSUP}, \quad (1)$$

Here, the ' λ ' represents the hyper-parameter used to fine-tune the impact of the unsupervised loss. Throughout the training process, the teacher model parameters ' θ_t ' are adapted from the student model parameters ' θ_s ' through the application of exponential moving average. This can be formulated as follows.

$$\theta_S \leftarrow \theta_S + \gamma \frac{\partial(\mathcal{L}_{SUP} + \lambda \mathcal{L}_{UNSUP})}{\partial \theta_S}, \theta_T \leftarrow \alpha \theta_T + (1 - \alpha) \theta_T \quad (2)$$

As can be seen, γ refers to the learning rate, and α to the EMA coefficient. Here EMA is applied to allow the teacher network to provide reliable pseudo-labels throughout training.

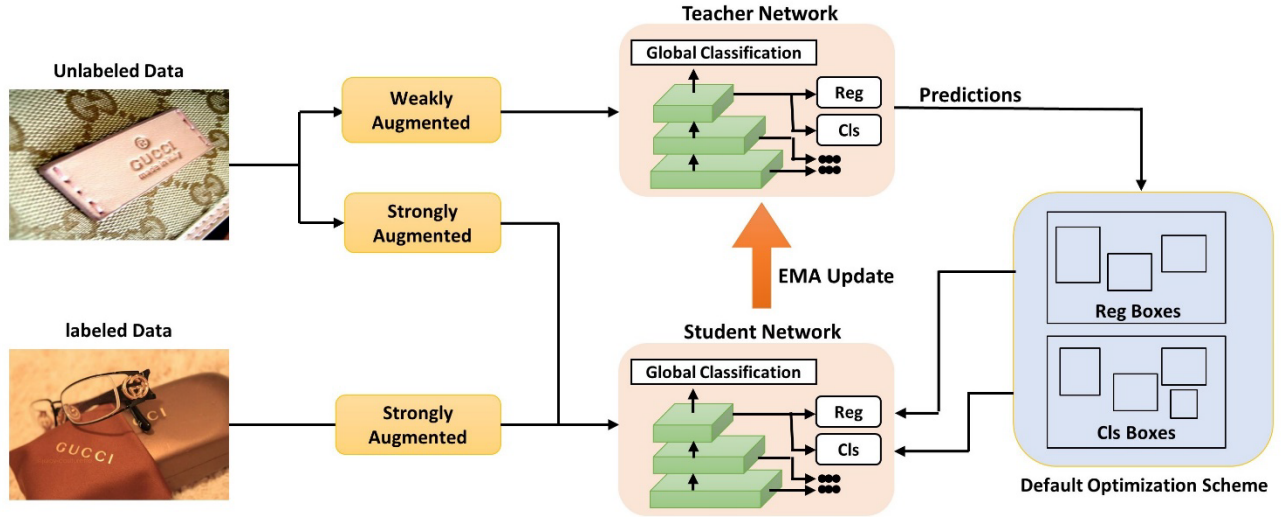


Fig. 1. The Proposed Architecture consist of Teacher Model which produces pseudo-labels for the student model. Its parameters are updated from the student model via EMA. The model uses a standard YOLOv8 Optimization scheme.

End-to-end detection is a common strategy applied by one-stage detection models, unlike two-stage networks, where assuming a labeled image will be $li \in r^{h \times w \times 3}$, while the single-stage detection model will give out a tensor $Ti \in r^{H \times W \times (C+5)}$ to handle the combined multitask prediction, where C stands for the count of object classes, 5 for the confidence score, and the bounding box coordinates. As a result, the supervised loss \mathcal{L}_{SUP} will be described as

$$\mathcal{L}_{SUP}(Ti, yi) = \mathcal{L}_{IOU}(Ti, yi^{COORD}) + \mathcal{L}_{CONF}(Ti, yi^{CONF}) \quad (3)$$

$$+ \mathcal{L}_{CLS}(Ti, yi^{CLS})$$

Here yi represents the labels for class, coordinates, and confidence score whereas binary cross entropy losses for confidence score regression and classification are \mathcal{L}_{CLS} and \mathcal{L}_{CONF} and IOU loss is denoted as \mathcal{L}_{IOU} . Concerning the unsupervised loss, a simple approach involves choosing predicted bounding boxes that surpass a predetermined threshold as pseudo-labels. This can be expressed as follows.

$$\mathcal{L}_{UNSUP}(T_U, y_U) = \mathcal{L}_{IOU}(T_U, y_U^{COORD}) + \mathcal{L}_{CONF}(T_U, y_U^{CONF}) + \mathcal{L}_{CLS}(T_U, y_U^{CLS}) \quad (4)$$

Here, the set of pseudo-labels is represented by y_U for object class, coordinates, and confidence.

IV. EXPERIMENT SETUP & RESULTS

A. Dataset and Evaluation Metrics

We trained the proposed model on the OpenLogo dataset [18]. We created sets of labelled images having percentages of 5%, 10%, and 15%, respectively, and utilized the other images as unlabelled data. We used mAP (mean average precision) at IoU threshold 50 (mAP50) and mAP at IoU threshold 50-95 (mAP50-95) as our evaluation metrics.

B. Model Training

We experimented with YOLOv8l and used it as a basic framework and started the training with random weights. We utilized the exponential moving average for momentarily

assembling the model parameters to further improve training stability. We additionally retained the data augmentations that were part of YOLOv8, such as the random horizontal flips, random image scaling, mosaic augmentation, and HSV color-space augmentation.

The hyperparameters configuration can be seen in TABLE 1. Furthermore, we adjusted several key hyperparameters in the teacher-student learning framework, particularly we lowered the pseudo-labeling threshold from 0.7, which is often set at higher

levels in two-stage methods, to 0.4. This tweak tackles the problem of noisy pseudo-labeling.

TABLE I. HYPERPARAMETER CONFIGURATION

<i>Hyperparameters</i>	<i>Settings</i>
Learning Rate	0.01
Momentum	0.937
EMA Coefficient	0.9996
Optimizer	SGD
Weight Decay	0.0005
Batch Size	16
Epochs	300

C. Results

Throughout the training process, we noticed that as we increased the volume of labelled data with the percentages of 5%, 10%, and 15%, mean Average Precision (mAP) kept increasing at a specific rate. However, after 200 epochs, a subsequent decline in mAP was observed. Overall, the

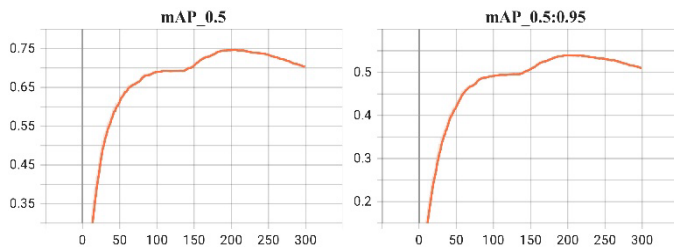


Fig. 2. Performance metric curve for the proposed method

mAP_0.5 peaked at 0.75 while mAP_0.5:0.95 peaked at 0.59 which can be seen in Fig 2. Training losses (box loss, cls loss, obj loss) kept relatively low at 0.012, 0.001, 0.005 respectively.

V. CONCLUSION

This paper presents a single-stage semi-supervised object detection approach, an area that has remained relatively unexplored in the existing literature. We introduced a student-teacher learning network tailored for the single-stage detection model. Furthermore, we have chosen the advanced YOLOv8 detection model as our basic framework. We have meticulously refined its implementation to fully leverage the advantages of semi-supervised object detection, and to assess the performance of our model, we employed the mean Average Precision (mAP) as our primary evaluation metric. The results obtained based on different settings support its effectiveness in addressing the identified primary two-stage semi-supervised object detection issues.

ACKNOWLEDGMENT

This work was funded by the Korea government (MSIT) (2022-0-00215), and this work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2021-0-02068, Artificial Intelligence Innovation Hub).

REFERENCES

- [1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 580–587, Nov. 2013, doi: 10.1109/CVPR.2014.81.
- [2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Trans Pattern Anal Mach Intell, vol. 39, no. 6, pp. 1137–1149, Jun. 2015, doi: 10.1109/TPAMI.2016.2577031.
- [3] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO Series in 2021," vol. 5, p. 12, Jul. 2021, Accessed: Sep. 19, 2023. [Online]. Available: <https://arxiv.org/abs/2107.08430v2>
- [4] G. Jocher et al., "ultralytics/yolov5: v3.0," Aug. 2020, doi: 10.5281/ZENODO.3983579.
- [5] K. Sohn, Z. Zhang, C.-L. Li, H. Zhang, C.-Y. Lee, and T. Pfister, "A Simple Semi-Supervised Learning Framework for Object Detection," May 2020, Accessed: Sep. 19, 2023. [Online]. Available: <https://arxiv.org/abs/2005.04757v2>
- [6] Y.-C. Liu et al., "Unbiased Teacher for Semi-Supervised Object Detection," Feb. 2021, Accessed: Sep. 19, 2023. [Online]. Available: <https://arxiv.org/abs/2102.09480v1>
- [7] Y. Tang, W. Chen, Y. Luo, and Y. Zhang, "Humble Teachers Teach Better Students for Semi-Supervised Object Detection," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3131–3140, Jun. 2021, doi: 10.1109/CVPR46437.2021.00315.
- [8] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," Adv Neural Inf Process Syst, vol. 2017-December, pp. 1196–1205, Mar. 2017, Accessed: Sep. 19, 2023. [Online]. Available: <https://arxiv.org/abs/1703.01780v6>
- [9] D. Berthelot, N. Carlini, I. Goodfellow, A. Oliver, N. Papernot, and C. Raffel, "MixMatch: A Holistic Approach to Semi-Supervised Learning," Adv Neural Inf Process Syst, vol. 32, May 2019, Accessed: Sep. 19, 2023. [Online]. Available: <https://arxiv.org/abs/1905.02249v2>
- [10] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning Augmentation Strategies From Data," pp. 113–123, 2019. Accessed: Sep. 19, 2023. [Online]. Available: <https://pillow.readthedocs.io/en/5.1.x/>
- [11] Q. Xie, Z. Dai, E. Hovy, M. T. Luong, and Q. V. Le, "Unsupervised Data Augmentation for Consistency Training," Adv Neural Inf Process Syst, vol. 2020-December, Apr. 2019, Accessed: Sep. 19, 2023. [Online]. Available: <https://arxiv.org/abs/1904.12848v6>
- [12] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," Apr. 2020, Accessed: Sep. 19, 2023. [Online]. Available: <https://arxiv.org/abs/2004.10934v1>
- [13] W. Liu et al., "SSD: Single Shot MultiBox Detector," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9905 LNCS, pp. 21–37, Dec. 2015, doi: 10.1007/978-3-319-46448-0_2.
- [14] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," Proceedings of the IEEE International Conference on Computer Vision, vol. 2017-October, pp. 2980–2988, Dec. 2017, doi: 10.1109/ICCV.2017.322.
- [15] "YOLOv8 - Ultralytics YOLOv8 Docs." <https://docs.ultralytics.com/models/yolov8/> (accessed Sep. 20, 2023).
- [16] E. Arazo, Di. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning," Proceedings of the International Joint Conference on Neural Networks, Aug. 2019, doi: 10.1109/IJCNN48605.2020.9207304.
- [17] D. Berthelot et al., "ReMixMatch: Semi-Supervised Learning with Distribution Matching and Augmentation Anchoring," 2020
- [18] "QMUL-OpenLogo." Accessed: Sep. 25, 2023. [Online]. Available: <https://hangsu0730.github.io/qmul-openlogo/>