Improving Semi-Supervised Object Detection by ROI-Enhanced Contrastive Learning

Teng-Kuan Huang and Mei-Chen Yeh
Department of Computer Science and Information Engineering,
National Taiwan Normal University, Taipei, Taiwan
E-mail: kensai1402@yahoo.com.tw, myeh@ntnu.edu.tw

***Abstract*—Semi-supervised object detection has emerged as a promising paradigm to alleviate the data annotation burden by utilizing a small set of labeled data in conjunction with a larger pool of unlabeled data. Current state-of-the-art methods commonly employ self-training strategies, using pseudo labels to learn from unlabeled data. However, pseudo labels are inherently noisy, particularly in the early stages of training. In this paper, we propose a contrastive learning approach to enhance semi-supervised object detection. Departing from conventional box-level predictions, our method introduces consistency regularization at the feature-level representation. Specifically, we leverage candidate boxes selected by the Region Proposal Network (RPN) for Region of Interest (RoI)-based contrastive learning and introduce pixel-level comparisons for spatial-aware loss calculation. Our experiments demonstrate that the proposed RoI-enhanced contrastive learning effectively enables the model to extract additional information from unlabeled data.**

I. Introduction

In recent years, remarkable progress in object detection has been driven by supervised learning methods, leveraging abundant labeled training data [1]. Yet, the resource-intensive process of obtaining accurate annotations has led to a growing interest in semi-supervised approaches. These approaches combine limited labeled data with a wealth of unlabeled data to optimize object detection performance [2], [3]. Within this domain, the Teacher-Student Mutual Learning framework, renowned for its effectiveness in self-training for semi-supervised object detection [4], has gained significant attention.

Despite the various methods for utilizing pseudo-labels generated by evolving teacher models, their universal applicability across diverse datasets remains a challenge, encompassing fixed confidence thresholds to sophisticated filtering mechanisms [5]–[7]. This paper proposes an alternative solution for leveraging unlabeled data in semi-supervised object detection.

The Teacher-Student framework introduces varying degrees of visual augmentation to unlabeled data, creating both strongly and weakly augmented image versions. Pseudo-labels are generated by the teacher model on weakly augmented images, while the student model learns from filtered pseudo-labels on strongly augmented counterparts. This process mirrors the contrastive learning approach commonly used in self-supervised learning, such as BYOL [8] and SimpleCL [9]. The distinction lies in the computation of the supervised loss or the similarity—where the former calculates the loss from prediction boxes, the latter does so from feature-level representations, as illustrated in Fig. 1.

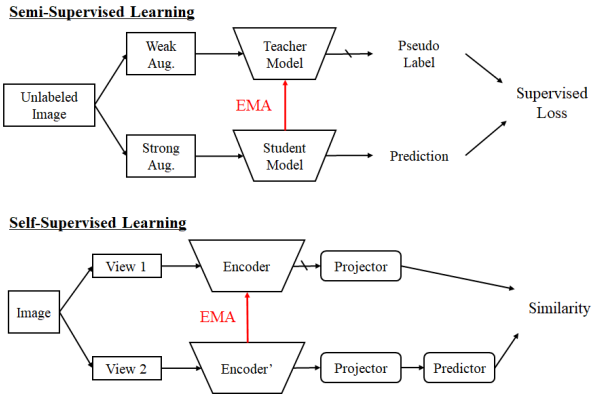


Fig. 1. Analogy between Semi-Supervised Learning and Self-Supervised Learning.

Despite the commonality in the self-training process, pseudo-labels generated by the teacher model inevitably introduce label noise. Nevertheless, deep neural networks have shown the ability to effectively memorize arbitrary noisy labels during training [10], emphasizing the importance of a strategy to enhance the utilization of unlabeled data. The current landscape of self-supervised learning methodologies has demonstrated the efficacy of learning visual representations from unlabeled data. Specifically, approaches grounded in instance discrimination, treating each image as a distinct class and utilizing a contrastive learning objective, have achieved significant success [8], [9], [11], [12]. In particular, contrastive learning seeks to encode image features by mapping multiple augmented views of the same image to a trainable embedding space. This label-free approach significantly diminishes the reliance on manual annotation in datasets, presenting a cost-effective solution for model training. While self-supervised learning methods exhibit success in downstream tasks even in the absence of a filtering process, our motivation lies in incorporating this strategy to foster feature learning independent of (pseudo) labels in the context of semi-supervised object detection.

However, when transitioning these learning methods to object detection datasets such as MS-COCO [13], a discernible decline in learning performance becomes evident. Most con-

trastive self-supervised methods are predicated on the assumption of semantic consistency in images. In these methods, the entire image is regarded as object-centric, assuming that a single instance predominantly occupies the image space, as is often the case in datasets like ImageNet [14]. While this assumption facilitates leveraging semantic consistency for effective representations in single-label classification tasks, real-world images deviate from this premise. In practice, images may comprise diverse semantic content, with instances varying in size and appearing at various locations within the image.

In this paper, we present a contrastive learning approach for semi-supervised object detection. Our approach involves processing images through the Faster-RCNN backbone [1] to generate feature maps, which are then subjected to the Region Proposal Network (RPN) to identify candidate boxes likely to be foreground objects. Unlike using image-level representation, we perform box-level contrastive learning by Region of Interest (RoI) pooling on feature maps generated from different augmented images at the same locations. Furthermore, we avoid global pooling during the contrastive loss calculation. Instead, we choose to compute the loss based on features extracted from corresponding locations in two augmentations. This spatially aware approach retains crucial feature information, as will be explored further in the ablation study (Section IV-C). The main contributions of this paper are summarized as follows:

- We present a contrastive learning approach to improve semi-supervised object detection, performing consistency regularization not only by aligning the box predictions to pseudo boxes but also by considering feature-level representations.
- To address the challenge of object detection, the contrastive loss is computed at the box level, rather than on the entire image. Furthermore, the loss computation is spatially aware.
- Through experiments, we demonstrate that contrastive learning on RoI features can enhance the model's ability to gain additional information from unlabeled data.

In the following sections, we begin by reviewing related studies in Section II. Section III outlines the main investigation method. Finally, experiments and results are discussed in Section IV, and the paper concludes in Section V.

II. Related Work

A. Semi-Supervised Learning

In the realm of semi-supervised learning, two predominant methodological approaches have emerged: one emphasizes data augmentation and perturbation of the input data, while the other centers around consistency regularization. Noteworthy examples include MixMatch [15], which applies multiple augmentations to unlabeled images, averages the predictions of these augmented images, and sharpens them as pseudo-labels for model training. Similarly, FixMatch [16] employs both strong and weak augmentations on unlabeled images, requiring the model's predictions to remain consistent across different versions of the same image. Both methodologies strive to maintain stable predictions across varied inputs, ensuring consistency even in the face of augmentation or perturbation [17].

However, the architectural intricacies of object detection models surpass those of image classification, demanding a more nuanced approach when adapting strategies from the image classification domain.

B. Semi-Supervised Object Detection

Recent advancements in semi-supervised object detection have been propelled by pseudo-labeling methods. STAC [3] initiates the process by generating pseudo-labels using a pre-trained model, subsequently fine-tuning the detection model by integrating these labels with strongly augmented data. Methods such as Unbiased Teacher [2], Instant Teaching [18], and Soft Teacher [19] dynamically generate pseudo-labels on weakly-augmented data while training the model on strongly-augmented data. This approach not only produces higher-quality pseudo-labels but also enhances the model's detection capabilities.

Pseudo-labeling strategies vary: Unbiased Teacher [2] employs a fixed confidence threshold to dynamically filter pseudo-labels, Active Teacher [7] utilizes three metrics derived from confidence values across the entire image, and Dense Learning [5] applies an Adaptive Filtering operation to segment pseudo-labels into foreground, background, and ignored areas. Label Matching [6] assumes that the class distribution in the unlabeled data should resemble that in the labeled data, adjusting thresholds dynamically to create high-quality pseudo-labels accordingly.

However, the generation of pseudo-labels inherently introduces label noise. Deep neural networks have demonstrated the ability to effectively learn from noisy labels during training [10], emphasizing the need for robust strategies to utilize unlabeled data in the context of semi-supervised object detection.

C. Self-Supervised Learning

Self-supervised learning has emerged as a compelling research area within computer vision, showcasing notable achievements in recent years. Diverging from traditional supervised learning, self-supervised approaches eliminate the need for manual data labeling. Instead, these methods capitalize on inherent data structures, providing implicit supervision through pretext tasks designed to learn meaningful representations. Pioneering works such as RotNet [20], jigsaw puzzles [21], and predicting relative patch positions [22] have laid the foundation for self-supervised learning, producing transferable embeddings that exhibit efficacy in tasks like image classification.

In recent advancements, contrastive learning has played a pivotal role. Methods like SimCLR [9] transform images into multiple views, minimizing the distance between identical views and maximizing distances between different views in the feature map. SimCLR, for instance, employs the InfoNCE loss [23] to align similar views in the embedding space. Distinctively, BYOL [8] utilizes asymmetric neural networks, eliminating the need for negative pairs during training. SwAV

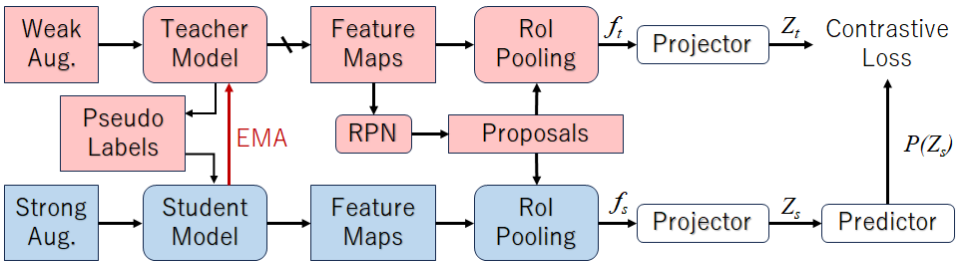


Fig. 2. Flowchart illustrating our semi-supervised object detection framework trained with RoI-enhanced contrastive learning.

[12] introduces an online clustering component to prevent training collapse, leveraging soft encoding and prototype-based predictions, all achieved without relying on a memory bank [11].

In summary, both contrastive learning in self-supervised fields and semi-supervised learning share the common goal of making consistent predictions across different augmented images. However, the distinction lies in the calculation of loss: self-supervised learning assesses feature similarity, while semi-supervised learning employs model-predicted labels to calculate loss. This key difference is underscored by the presence of a small amount of labeled data in semi-supervised learning, enabling the use of meaningful pseudo-labels in subsequent training. Incorrect pseudo-labels can lead to the model learning erroneous information, particularly in semi-supervised object detection. Thus, effective strategies for pseudo-label filtering and generation become crucial research directions. Combining the feature-based learning concept from self-supervised learning could offer a novel direction for pseudo-labeling methods, enabling the model to extract more valuable information from unlabeled data while striking a balance between the quality and quantity of pseudo-labels.

III. Method

In this section, we begin by outlining the semi-supervised object detection problem. Subsequently, we introduce our proposed semi-supervised object detection model, which leverages two approaches for learning from unlabeled data: one based on pseudo labels and another based on box-level feature representations.

A. Problem Definition

The semi-supervised object detection problem involves a set of labeled data $D_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ and a set of unlabeled data $D_u = \{x_i^u\}_{i=1}^{N_u}$ used for training the model, where N_s and N_u represent the numbers of labeled and unlabeled images, respectively. Each labeled image x^s is accompanied by corresponding annotations y^s, encompassing the positions, dimensions, and classes of bounding boxes capturing foreground objects.

B. Semi-Supervised Object Detection with Contrastive Learning

1) Background: Our object detection model builds upon Faster-RCNN [1], incorporating a Feature Pyramid Network (FPN) and ResNet-50 [24]. Initially trained with a limited set of labeled data (e.g., 1% of the COCO training set), the model optimizes the supervised loss L_{sup} involving box-regression and box-classification losses——a stage referred to as the burn-in.

Following the Unbiased Teacher approach [2], the model transitions to the teacher-student mutual learning stage. Duplicating the detector generates two models (teacher θ_t and student θ_s). In this stage, the teacher model generates pseudo-labels on weakly-augmented images, filtered through a confidence threshold. The student model then employs these filtered pseudo-labels for supervised training on strongly-augmented images, computing the object detection loss L_{unsup}. This iterative process creates a symbiotic relationship, with the teacher guiding the student and the student enhancing the quality of pseudo-labels.

To address potential inaccuracies in pseudo-labels, especially during early training, we propose leveraging feature-level representations from the feature maps to enhance the utilization of unlabeled data.

2) RoI-Enhanced Contrastive Learning: We extend Unbiased Teacher with contrastive learning, introducing two projectors and one predictor (Fig. 2). After generating proposals, we perform RoI pooling on the feature maps of both teacher and student models. A projector is added to the teacher branch, and a projector and a predictor are added to the student branch. Exploiting the inherent strong and weak augmentations, we compute additional contrastive learning loss L_{CL}. Utilizing Faster-RCNN, we focus contrastive learning on positions selected through the Region Proposal Network (RPN), likely to be foreground objects. Specifically, we consider only the top 10 proposals based on the likelihood of being in the foreground returned from the teacher RPN.

Given these top 10 proposals, we perform RoI pooling on the student model's feature map, obtaining corresponding RoI features f_t and f_s. We then feed f_t to the projector to obtain Z_t, and feed f_s to the projector and the predictor to obtain Z_s. Utilizing an asymmetric network for contrastive learning [8], which enhances feature-level representations, we compute the contrastive loss as follows:

$$L_{CL} = 2 - 2 \cdot \frac{\langle Z_t, P(Z_s) \rangle}{\|Z_t\|_2 \cdot \|P(Z_s)\|_2}. \quad (1)$$

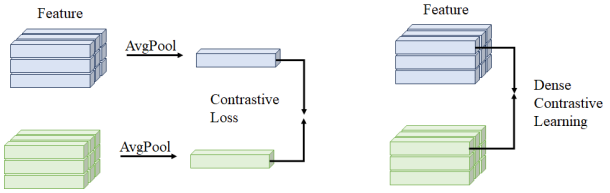


Fig. 3. Difference between the spatial-unaware (left) and the spatial-aware (right) loss calculation.

3) Dense Contrastive Learning: Contrary to applying global pooling to RoI features for contrastive loss calculation, we retain spatial information in the RoI features. We compute the contrastive loss across all positions in both teacher and student model features, as in [25]. Figure 3 illustrates the spatial-aware loss calculation.

The projector and predictor are implemented as multiple-layer perceptrons (MLPs), following [25], [26]. All features entering the projector avoid global pooling, with the projector consisting of three 1×1 convolutional layers. Each layer, except the final one, is followed by a batch normalization layer and a ReLU layer. The predictor comprises two 1×1 convolutional layers, with the first having an output dimension of 64, followed by batch normalization and ReLU, and the second having an output dimension of 256.

Finally, the parameters of the student model are updated through the supervised, unsupervised detection and contrastive loss:

$$L_{total} = L_{sup} + L_{unsup} + L_{CL}. \tag{2}$$

The exponential moving average (EMA) strategy updates the teacher model from the trained student model for each iteration:

$$\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s. \tag{3}$$

During inference, the teacher model is utilized for predictions.

IV. Experiments

A. Experimental Settings

This section presents the results of our experiments. To ensure a fair comparison across methods, we adopted Faster-RCNN with ResNet50-FPN as the backbone for object detection, following the setup in [2]. A confidence threshold of $\delta = 0.7$ was utilized. Data augmentation techniques included random horizontal flips for weak augmentation, along with random color jittering, grayscale, Gaussian blur, and cut-out patches for strong augmentations. The exponential moving average (EMA) rate α was set to 0.99. However, due to computational resource constraints, our model was trained with a smaller batch size (16).

We conducted our experiments on the MS-COCO dataset [13]. In line with [2], we randomly sampled 1%, 5%, and 10% of images from the approximately 118,000 images in the `train2017` subset to serve as labeled training data. The remaining images were utilized as unlabeled data. The test set comprised 5,000 images from `val2017`. Evaluation metrics were based on $AP_{50:95}$, denoted as mAP, and the performance was assessed on the teacher model.

B. Results

Our proposed method is designed as a versatile, plug-and-play framework compatible with various teacher-student architectures. Developed initially based on the Unbiased Teacher [2] framework, we exclusively compare our method with Unbiased Teacher in this study to showcase the adaptability within this well-established teacher-student paradigm.

Table I displays the detection performance of the supervised baseline, Unbiased Teacher [2], and our method. Our proposed method outperformed Unbiased Teacher across all settings, achieving improvements of 2.93%, 1.89%, and 1.01% for the 1%, 5%, and 10% labeled data settings, respectively. The mAPs at different iterations are visualized in Fig. 4. It's important to note that both methods underwent the same burn-in process.

TABLE I
Experimental results on COCO-standard.

Method	1%	5%	10%
Supervised	9.05	18.47	23.86
Unbiased Teacher [2]	20.19	28.20	31.46
Ours	**20.78**	**28.73**	**31.77**

C. Ablation Study

To gain a deeper understanding of the proposed method, we conducted additional experiments focusing on the impact of RoI-enhanced contrastive learning, spatial-aware loss computation, and the selection of the top-scored RoI. In these experiments, 10% of the images from the training set were used to train the models, and mAP values were computed on COCO `val2017`. To examine the effect of different contrastive learning choices, we disabled the computation of consistency between the teacher and the student models based on pseudo-labels and focused solely on feature-level consistency learned from unlabeled data.

1) Effect of RoI-Enhanced Contrastive Learning: In contrastive learning, we considered features from the candidate boxes returned by RPN, rather than directly using features obtained from the backbone network. We hypothesized that, during the training of the Faster R-CNN-based object detector, the RPN also learns to discover foreground objects. Intersection over Union (IoU) was calculated between the candidate boxes and the ground truth boxes to categorize all boxes into highly overlapped (positive) or lowly overlapped/non-overlapped (negative) samples. Confusing candidate boxes that did not sufficiently overlap with ground truth were excluded from loss calculation. This approach ensures that the contrastive learning process focuses exclusively on object-level information. Figure 5 (a) illustrates the performance comparison of using RoI features versus the entire feature map. RoI-based features consistently outperformed the alternative after reaching 20,000 iterations, validating the effectiveness of RoI-based contrastive learning.

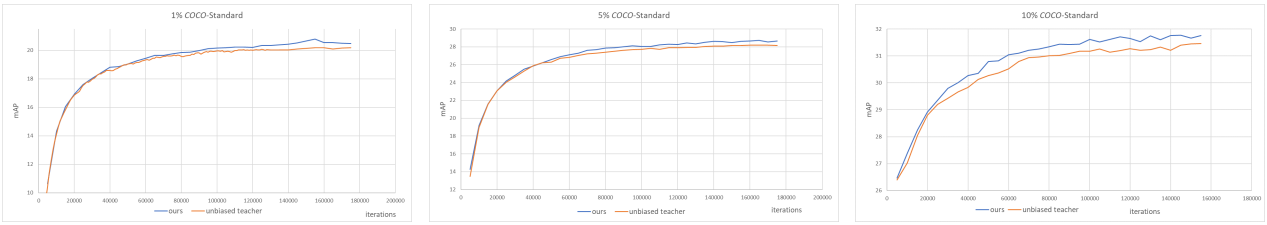
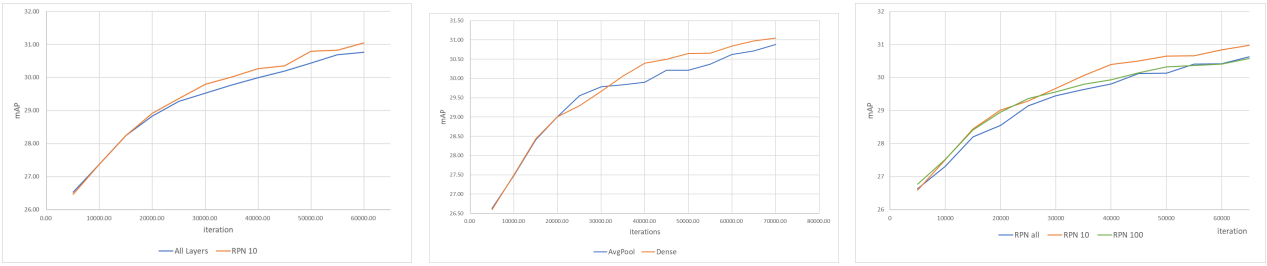


Fig. 4. Comparison with Unbiased Teacher [2] on COCO-Standard using 1%, 5%, and 10% labeled training data, showcasing the mAP values at different training iterations.



(a) (b) (c)

Fig. 5. Performance comparison of (a) contrastive learning with RoI features and the entire feature map; (b) spatial-unaware and spatial-aware contrastive learning; (c) using different numbers of RoIs.

2) Effect of AvgPool: When computing the contrastive loss, we consider that if the features extracted from the candidate boxes undergo global pooling, as in conventional methods, spatial information would be lost, potentially impacting the true similarity between two boxes. Figure 5 (b) displays the performance comparison of spatial-unaware (using global pooling on the features) and the proposed spatial-aware methods. Our approach consistently achieved higher mAP values after 30,000 iterations.

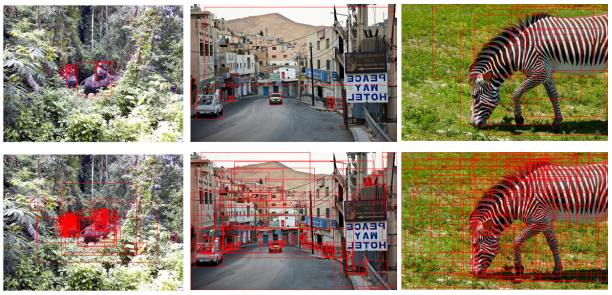


Fig. 6. Visualization of top 10 (top row) and top 100 (bottom row) high-scoring RoIs.

3) Number of RPN Proposals: In the proposed method, we utilized the top 10 boxes in contrastive learning. Alternatively, we experimented with other choices, including using the top 100 and all candidate boxes. Figure 5 (c) shows the comparison result. In this experiment, Non-Maximum Suppression (NMS) with an IoU threshold of 0.5 was used to eliminate overlapping boxes. After NMS, each image retained around 300 candidate boxes out of the initial 1000 boxes. Figure **??** illustrates that when using all candidate boxes, the model's mAP is generally lowest. When using top 100 candidate boxes, the model's early-stage mAP is roughly comparable to that of using the top 10; however, it falls behind after 30,000 iterations. To further investigate, we randomly selected three samples and visualized the RoIs in Figure 6, revealing that many boxes in the top 100 selection set do not effectively capture objects. Consequently, the effectiveness of performing contrastive learning on these boxes is questionable. Therefore, we adopted the top 10 high-scoring candidate boxes for contrastive learning.

V. Conclusion

In this paper, we have presented a contrastive learning approach to enhance semi-supervised object detection. Specifically, we leverage the candidate boxes selected by the Region Proposal Network (RPN) to facilitate RoI-based contrastive learning. Additionally, we incorporate pixel-level comparisons to enable spatial-aware loss calculation. Our experimental results showcase the effectiveness of the proposed RoI-enhanced contrastive learning, coupled with pseudo-labeling, in extracting valuable information from unlabeled data. This integration not only enhances model performance but also mitigates the impact of noisy pseudo-labels during consistency regularization in semi-supervised object detection.

One future direction we have been pursuing is the validation of the proposed plug-and-play method on alternative detection frameworks beyond Faster-RCNN [1]. Expanding our approach to different models will provide valuable insights into its adaptability and generalizability across various architectures. By extending our methodology to a broader spectrum of detectors, we aim to offer a robust and flexible solution that can be tailored to different detection paradigms,

ultimately advancing the state-of-the-art in semi-supervised object detection.

ACKNOWLEDGMENT

This work was supported by the National Science and Technology Council of Taiwan (MOST 111-2221-E-003-016-MY2, MOST 110-2634-F-002-050).

REFERENCES

[1] R. Girshick, "Fast r-cnn," in *IEEE International Conference on Computer Vision*, 2015.

[2] Y.-C. Liu *et al.*, "Unbiased teacher for semi-supervised object detection," *arXiv preprint arXiv:2102.09480*, 2021.

[3] K. Sohn *et al.*, "A simple semi-supervised learning framework for object detection," *arXiv preprint arXiv:2005.04757*, 2020.

[4] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[5] B. Chen *et al.*, "Dense learning based semi-supervised object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[6] B. Chen *et al.*, "Label matching semi-supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[7] P. Mi *et al.*, "Active teacher for semi-supervised object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[8] J.-B. Grill *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, pp. 21 271–21 284, 2020.

[9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, PMLR, 2020.

[10] D. Arpit, S. Jastrzebski, N. Ballas, *et al.*, "A closer look at memorization in deep networks," in *International Conference on Machine Learning*, PMLR, 2017.

[11] K. He *et al.*, "Momentum contrast for unsupervised visual representation learning," in *IEEE/CVF International Conference on Computer Vision*, 2020.

[12] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *arXiv:2006.09882*, 2020.

[13] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014.

[14] O. Russakovsky, J. Deng, H. Su, *et al.*, "Imagenet large scale visual recognition challenge," in *arXiv:1409.0575*, 2014.

[15] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Advances in Neural Information Processing Systems*, 2019.

[16] K. Sohn *et al.*, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in Neural Information Processing Systems*, 2020.

[17] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in Neural Information Processing Systems*, 2017.

[18] Q. Zhou *et al.*, "Instant-teaching: An end-to-end semi-supervised object detection framework," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[19] M. Xu *et al.*, "End-to-end semi-supervised object detection with soft teacher," in *IEEE/CVF International Conference on Computer Vision*, 2021.

[20] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *International Conference on Learning Representations*, 2018.

[21] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European Conference on Computer Vision*, 2016.

[22] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015.

[23] A. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," in *arXiv:1807.03748*, 2018.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[25] X. Wang *et al.*, "Dense contrastive learning for self-supervised visual pre-training," in *IEEE/CVF International Conference on Computer Vision*, 2021.

[26] X. Chen and K. He, "Exploring simple siamese representation learning," in *IEEE/CVF International Conference on Computer Vision*, 2021.