Semi-DETR: Semi-Supervised Object Detection with Detection Transformers

Jiacheng Zhang[1,2*] Xiangru Lin[2*] Wei Zhang[2] Kuo Wang[1] Xiao Tan[2]
Junyu Han[2] Errui Ding[2] Jingdong Wang[2] Guanbin Li[1,3†]
[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]Department of Computer Vision Technology (VIS), Baidu Inc., China
[3]Research Institute, Sun Yat-sen University, Shenzhen, China
{zhangjch58, wangk229}@mail2.sysu.edu.cn, liguanbin@mail.sysu.edu.cn
{linxiangru,zhangwei99,tanxiao01,hanjunyu,dingerrui,wangjingdong}@baidu.com

Abstract

We analyze the DETR-based framework on semi-supervised object detection (SSOD) and observe that (1) the one-to-one assignment strategy generates incorrect matching when the pseudo ground-truth bounding box is inaccurate, leading to training inefficiency; (2) DETR-based detectors lack deterministic correspondence between the input query and its prediction output, which hinders the applicability of the consistency-based regularization widely used in current SSOD methods. We present Semi-DETR, the first transformer-based end-to-end semi-supervised object detector, to tackle these problems. Specifically, we propose a Stage-wise Hybrid Matching strategy that combines the one-to-many assignment and one-to-one assignment strategies to improve the training efficiency of the first stage and thus provide high-quality pseudo labels for the training of the second stage. Besides, we introduce a Cross-view Query Consistency method to learn the semantic feature invariance of object queries from different views while avoiding the need to find deterministic query correspondence. Furthermore, we propose a Cost-based Pseudo Label Mining module to dynamically mine more pseudo boxes based on the matching cost of pseudo ground truth bounding boxes for consistency training. Extensive experiments on all SSOD settings of both COCO and Pascal VOC benchmark datasets show that our Semi-DETR method outperforms all state-of-the-art methods by clear margins.

1. Introduction

Semi-supervised object detection (SSOD) aims to boost the performance of a fully-supervised object detector by exploiting a large amount of unlabeled data. Current state-of-the-art SSOD methods are primarily based on object detectors with many hand-crafted components, e.g., rule-based label assigner [9, 26, 27, 31] and non-maximum suppression (NMS) [1] post-processing. We term this type of object detector as a traditional object detector. Recently, DETR [2], a simple transformer-based end-to-end object detector, has received growing attention. Generally, the DETR-based framework builds upon transformer [32] encoder-decoder architecture and generates unique predictions by enforcing a set-based global loss via bipartite matching during training. It eliminates the need for various hand-crafted components, achieving state-of-the-art performance in fully-supervised object detection. Although the performance is desirable, how to design a feasible DETR-based SSOD framework remains under-explored. There are still no systematic ways to fulfill this research gap.

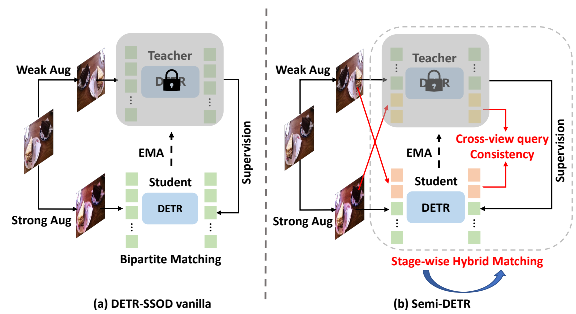


Figure 1. Comparisons between the vanilla DETR-SSOD framework based on the Teacher-Student architecture and our proposed Semi-DETR framework. Semi-DETR consists of Stage-wise Hybrid Matching, Cross-view Query Consistency powered by a cost-based pseudo label mining strategy.

Designing an SSOD framework for DETR-based detectors is non-trivial. Concretely, DETR-based detectors take a one-to-one assignment strategy where the bipartite-matching algorithm forces each ground-truth (GT) bound-

*Equally-contributed authors. Work done during an internship at Baidu.
†Corresponding author.

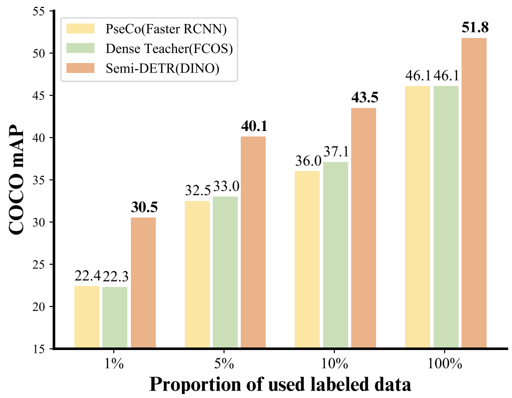


Figure 2. Performance comparisons between the proposed Semi-DETR and other SSOD methods, including PseCo [16] and Dense-Teacher [42].

ing box to match a candidate proposal as positive, treating remains as negatives. It goes well when the ground-truth bounding boxes are accurate. However, directly integrating DETR-based framework with SSOD is problematic, as illustrated in Fig. 1 (a) where a DETR-SSOD vanilla framework utilizes DETR-based detectors to perform pseudo labeling on unlabeled images. In the Teacher-Student architecture, the teacher model usually generates noisy pseudo bounding boxes on the unlabeled images. When the pseudo bounding box is inaccurate, the one-to-one assignment strategy is doomed to match a single inaccurate proposal as positive, leaving all other potential correct proposals as negative, thus causing learning inefficiency. As a comparison, the one-to-many assignment strategy adopted in the traditional object detectors maintains a set of positive proposals, having a higher chance of containing the correct positive proposal. On the one hand, the one-to-one assignment strategy enjoys the merits of NMS-free end-to-end detection but suffers the training inefficiency under semi-supervised scenarios; on the other hand, the one-to-many assignment strategy obtains candidate proposal set with better quality making the detector optimized more efficiently but inevitably resulted in duplicate predictions. Designing a DETR-based SSOD framework that embraces these two merits could bring the performance to the next level.

Additionally, the consistency-based regularization commonly used in current SSOD methods becomes infeasible in DETR-based SSOD. Specifically, current SSOD methods [3, 10, 13, 16] utilize consistency-based regularization to help object detectors learn potential feature invariance by imposing consistency constraints on the outputs of pairs-wise inputs (such as scale consistency [3, 10, 16], weak-strong consistency [13], etc.). Since the input features are deterministic in traditional object detectors, there is a one-to-one correspondence between the inputs and outputs, which makes the consistency constraint convenient to implement. However, this is not the case in DETR-based detectors. DETR-based detectors [2, 15, 20, 40, 44] use randomly initialized learnable object queries as inputs and constantly update the query features through the attention mechanism. As the query features update, the corresponding prediction results constantly change, which has been verified in [15]. In other words, there is no deterministic correspondence between the input object queries and its output prediction results, which prevents consistency regularization from being applied to DETR-based detectors.

According to the above analysis, we propose a new DETR-based SSOD framework based on the Teacher-Student architecture, which we term Semi-DETR presented in Fig. 1 (b). Concretely, we propose a **Stage-wise Hybrid Matching** module that imposes two stages of training using the one-to-many assignment and the one-to-one assignment, respectively. The first stage aims to improve the training efficiency via the one-to-many assignment strategy and thus provide high-quality pseudo labels for the second stage of one-to-one assignment training. Besides, we introduce a **Cross-view Query Consistency** module that constructs cross-view object queries to eliminate the requirement of finding deterministic correspondence of object queries and aids the detector in learning semantically invariant characteristics of object queries between two augmented views. Furthermore, we devise a **Cost-based Pseudo Label Mining** module based on the Gaussian Mixture Model (GMM) that dynamically mines reliable pseudo boxes for consistency learning according to their matching cost distribution. Differently, Semi-DETR is tailored for DETR-based framework, which achieves new SOTA performance compared to the previous best SSOD methods as illustrated in Fig. 2.

To sum up, this paper has the following contributions:

- We present a new DETR-based SSOD method based on the Teacher-Student architecture, called Semi-DETR. To our best knowledge, we are the first to examine the DETR-based detectors on SSOD, and we identify core issues in integrating DETR-based detectors with the SSOD framework.
- We propose a stage-wise hybrid matching method that combines the one-to-many assignment and one-to-one assignment strategies to address the training inefficiency caused by the inherent one-to-one assignment within DETR-based detectors when applied to SSOD.
- We introduce a consistency-based regularization scheme and a cost-based pseudo-label mining algorithm for DETR-based detectors to help learn semantic feature invariance of object queries from different augmented views.
- Extensive experiments show that our Semi-DETR method outperforms all previous state-of-the-art methods by clear margins under various SSOD settings on both MS COCO and Pascal VOC benchmark datasets.

2. Related Work

Semi-Supervised Object Detection. In SSOD, Pseudo Labeling [21, 25, 29, 33, 34, 36–38, 43] and Consistency-based Regularization [3, 10, 13, 16, 22, 24, 27, 42] are two commonly used strategies. A detailed description can be found in the supplementary document. However, most of these works are based on the traditional detectors, e.g. Faster RCNN [27], which involves many hand-crafted components, e.g anchor box, NMS, etc. Our Semi-DETR is significantly different from previous works: (1) we explored the challenges of the DETR-based object detectors on SSOD, which, to our best knowledge, is the first systematic research endeavor in SSOD; (2) our Semi-DETR method is tailored for the DETR-based detectors, which eliminates the training efficiency caused by bipartite matching with the noisy pseudo labels and presents a new consistency scheme for set-based detectors.

End-to-End Object Detection with Transformer. The pioneering work DETR [2] introduced transformers into object detection to eliminate the need for complex hand-crafted components in traditional object detectors. Many follow-up works have been dedicated to solving the slow convergence and high complexity issues of DETR [15, 20, 23, 39, 40, 44]. Recently, DINO [40] combined with a variety of improvements related to DETR, such as query selection [39, 44], contrastive query denoising [15], and achieved SOTA performance across various object detection benchmark datasets with excellent convergence speed. Complementary to these, we aim to extend the study of DETR-based detectors to SSOD and present Semi-DETR, which is a tailored design for SSOD. Our framework is agnostic to the choice of DETR-based detectors and could easily integrate with all DETR-based detectors. Omni-DETR [35] is a DETR-based object detector designed for omni-supervised object detection. It is not designed specifically for SSOD as admitted in their paper, but it is extended to the SSOD task by introducing a simple pseudo-label filtering scheme. Our Semi-DETR is significantly different from Omni-DETR in the following aspects: (1) Different motivations for model design; (2) Different training strategy; (3) Significant performance improvement. The detailed discussion is in Supplementary Document.

3. Semi-DETR

3.1. Preliminary

We aim to address the problem of semi-supervised DETR-based object detection, where a labeled image set $D_s = \{x_i^s, y_i^s\}_{i=1}^{N_s}$ and an unlabeled image set $D_u = \{x_i^u\}_{i=1}^{N_u}$ are available during training. N_s and N_u denote the amount of labeled and unlabeled images. For the labeled images x^s, the annotations y^s contain the coordinates and object categories of all bounding boxes.

3.2. Overview

The overall framework of our proposed Semi-DETR is illustrated in Fig. 3. Following the popular teacher-student paradigm [30] for SSOD, our proposed Semi-DETR adopts a pair of teacher and student models with exactly the same network architecture. Here we adopt DINO [40] as an example while the overall framework of Semi-DETR is compatible with other DETR-based detectors. Specifically, in each training iteration, weak-augmented and strong-augmented unlabeled images are fed to the teacher and student, respectively. Then the pseudo labels generated by the teacher with confidence scores larger than τ_s served as supervisions for training the student. The parameters of the student are updated by back-propagation, while the parameters of the teacher model are the exponential moving average (EMA) of the student. Our main contribution contains three new components: stage-wise hybrid matching, cross-view query consistency, and cost-based pseudo label mining, which address the core issues of DETR-based SSOD. In the following sections, we introduce more details of our proposed Semi-DETR.

3.3. Stage-wise Hybrid Matching

DETR-based frameworks rely on one-to-one assignment for end-to-end object detection. For DETR-based SSOD framework, an optimal one-to-one assignment $\hat{\sigma}_{o2o}$ can be obtained by performing the Hungarian algorithm between the predictions of the student and pseudo-labels generated by the teacher:

$$\hat{\sigma}_{o2o} = \underset{\sigma \in \xi_N}{\arg\min} \sum_{i=1}^{N} \mathcal{C}_{\text{match}}\left(\hat{y}_i^t, \hat{y}_{\sigma(i)}^s\right) \tag{1}$$

where ξ_N is the set of permutations of N elements and $\mathcal{C}_{\text{match}}\left(\hat{y}_i^t, \hat{y}_{\sigma(i)}^s\right)$ is the matching cost between the pseudo-labels $\hat{y}^t$ and the prediction of the student with index $\sigma(i)$.

However, in the early stage of SSOD training, the pseudo-labels generated by the teacher are usually inaccurate and unreliable, which imposes a high risk of generating sparse and low-quality proposals under the one-to-one assignment strategy. To exploit multiple positive queries to realize efficient semi-supervised learning, we propose to replace the one-to-one assignment with the one-to-many assignment:

$$\hat{\sigma}_{o2m} = \left\{ \underset{\boldsymbol{\sigma_i} \in C_N^M}{\arg\min} \sum_{j=1}^{M} \mathcal{C}_{\text{match}}\left(\hat{y}_i^t, \hat{y}_{\boldsymbol{\sigma_i}(j)}^s\right) \right\}_{i=1}^{|\hat{y}^t|}. \tag{2}$$

where C_N^M is the combination of M and N, which denotes that a subset of M proposals is assigned to each pseudo box $\hat{y}_i^t$. Following [6, 7], we utilize a high-order combination of

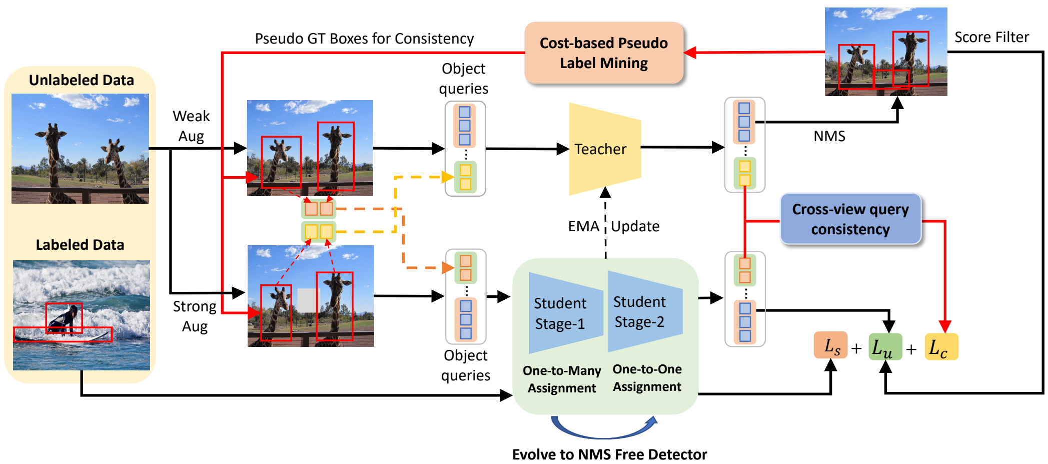


Figure 3. Overview of Semi-DETR. Our framework is based on Teacher-Student architecture. Specifically, a multi-stage training strategy is derived to avoid incorrect bipartite matching with low-quality pseudo labels. Hybrid Matching with one-to-many assignment is applied in the first stage to generate higher quality pseudo labels for the following one-to-one training stage. Besides, a cross-view query consistency loss is designed to further enhance the consistency learning for the whole training process, where the pseudo boxes are filtered by a cost-based GMM mining module.

classification score s and the IoU value u as the matching cost metric:

$$m = s^{\alpha} \cdot u^{\beta} \tag{3}$$

where α and β control the effect of classification score and IoU during the assignment, and following [6], we set $\alpha = 1$, $\beta = 6$ by default. With the one-to-many assignment, M proposals with the largest m values are selected as positive samples while regarding the remaining proposals as negative ones.

We train the model with one-to-many assignment for T_1 iterations in the early stage of semi-supervised training. Following [6, 17], the classification loss and regression loss are also modified at this stage:

$$L_{cls}^{o2m} = \sum_{i=1}^{N_{\text{pos}}} |\hat{m}_i - s_i|^{\gamma}\, BCE\left(s_i, \hat{m}_i\right) + \sum_{j=1}^{N_{\text{neg}}} s_j^{\gamma} BCE\left(s_j, 0\right) \tag{4}$$

$$L_{reg}^{o2m} = \sum_{i=1}^{N_{\text{pos}}} \hat{m}_i L_{GIoU}\left(b_i, \hat{b}_i\right) + \sum_{i=1}^{N_{\text{pos}}} \hat{m}_i L_{L_1}\left(b_i, \hat{b}_i\right) \tag{5}$$

$$L^{o2m} = L_{cls}^{o2m} + L_{reg}^{o2m} \tag{6}$$

where γ is set to 2 by default. With multiple assigned positive proposals for each pseudo label, the potentially high-quality positive proposals also get the chance to be optimized, which greatly improves the convergence speed and, in turn, obtains pseudo labels with better quality. However, the multiple positive proposals for each pseudo label result in duplicate predictions. To mitigate this problem, we propose to switch back to the one-to-one assignment training in the second stage. By doing this, we enjoy the high-quality pseudo labels after the first stage training and gradually reduce duplicate predictions to reach an NMS-free detector with one-to-one assignment training at the second stage. The loss functions of this stage are the same as [40]:

$$L^{o2o} = L_{cls}^{o2o} + L_{reg}^{o2o} \tag{7}$$

3.4. Cross-view Query Consistency

Traditionally, in non-DETR-based SSOD frameworks, consistency regularization can be employed conveniently by minimizing the difference between the output of teacher f_{θ} and student f'_{θ}, given the same input x with different stochastic augmentation:

$$\mathcal{L}_c = \sum_{\mathbf{x} \in \mathcal{D}_u} \text{MSE}\left(f_{\theta}(\mathbf{x}), f'_{\theta}(\mathbf{x})\right) \tag{8}$$

However, for DETR-based frameworks, as there is no clear (or deterministic) correspondence between the input object queries and their output prediction results, conducting consistency regularization becomes infeasible. To overcome this issue, we propose a Cross-view Query Consistency module that enables the DETR-based framework to learn semantically invariant characteristics of object queries between different augmented views.

Fig. 4 illustrates our proposed cross-view query consistency module. Specifically, for each unlabeled image, given a set of pseudo bounding boxes b, we process the RoI features extracted via RoIAlign [11] with several MLPs:

$$\begin{aligned} c_t &= \text{MLP}(\text{RoIAlign}(F_t, b)) \\ c_s &= \text{MLP}(\text{RoIAlign}(F_s, b)) \end{aligned} \tag{9}$$

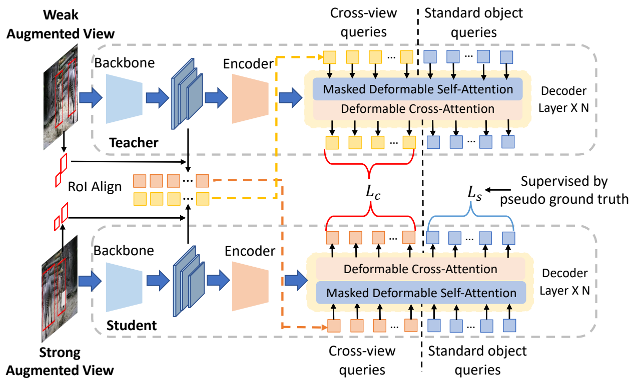


Figure 4. Overview of the cross-view query consistency module. Query embeddings from the RoI features of pseudo labels on different views are cross-wise sent to the teacher and student decoders. The corresponding decoded features are enforced to be similar by a consistency loss.

where F_t and F_s denote the backbone feature of the teacher and student, respectively. Subsequently, c_t and c_s are regarded as cross-view query embeddings and attached to the original object queries in another view to serve as the input of the decoder:

$$\begin{aligned} \hat{o}_t, o_t &= \text{Decoder}_t([c_s, q_t], E_t|A) \\ \hat{o}_s, o_s &= \text{Decoder}_s([c_t, q_s], E_s|A) \end{aligned} \tag{10}$$

where $q_.$ and $E_.$ denote the original object queries and the encoded image features, respectively. $\hat{o}_.$ and $o_.$ denote the decoded features of cross-view queries and original object queries. Note the subscript t and s indicate teacher and student, respectively. Following [15], the attention mask A is also employed to avoid information leakage.

With the semantic guide of input cross-view queries embeddings, the correspondence of the decoded features can be naturally guaranteed, and we impose consistency loss as follows:

$$\mathcal{L}_c = -\frac{c_t \cdot c_s}{\|c_t\| \times \|c_s\|} \tag{11}$$

3.5. Cost-based Pseudo Label Mining

To mine more pseudo boxes with meaningful semantic contents for the cross-view query consistency learning, we propose a cost-based pseudo label mining module that dynamically mines reliable pseudo boxes in the unlabeled data. Specifically, we perform an additional bipartite matching between the initial filtered pseudo boxes and the predicted proposals and utilize the matching cost to describe the reliability of the pseudo boxes:

$$C_{ij} = \lambda_1 C_{Cls}(p_i, \hat{p_j}) + \lambda_2 C_{GIoU}(b_i, \hat{b_j}) + \lambda_3 C_{L_1}(b_i, \hat{b_j}) \tag{12}$$

where p_i, b_i represents the classification and regression result of i-th predicted proposals while $\hat{p_j}$, $\hat{b_j}$ indicates the class label and box coordinates of j-th pseudo label.

Subsequently, in each training batch, we cluster the initial pseudo boxes into two states by fitting a Gaussian Mixture Model for the matching cost distribution. As illustrated in Figure 4, the matching cost aligns well with the quality of pseudo boxes. We further set the cost value of the clustering center of the reliable ones as the threshold and collect all pseudo boxes with lower cost than the threshold for the cross-view query consistency calculation.

3.6. Loss Function

The final loss L is represented as follows:

$$\begin{aligned} L = &\mathbb{I}(t \leq T_1) \cdot (L_{sup}^{o2m} + w_u \cdot L_{unsup}^{o2m}) \\ &+ \mathbb{I}(t{>}T_1) \cdot (L_{sup}^{o2o} + w_u \cdot L_{unsup}^{o2o}) \\ &+ w_c \cdot L_c \end{aligned} \tag{13}$$

where $L_{sup}^{\cdot}$ and $L_{unsup}^{\cdot}$ are the supervised loss and the unsupervised loss, respectively, containing both the classification loss and regression loss. The L_c means the cross-view consistency loss. The w_u and w_c are the unsupervised loss weight and consistency loss weight, which set $w_u = 4$ and $w_c = 1$ by default. t is the current training iteration and T_1 is the duration time of the first stage training within the SHM module.

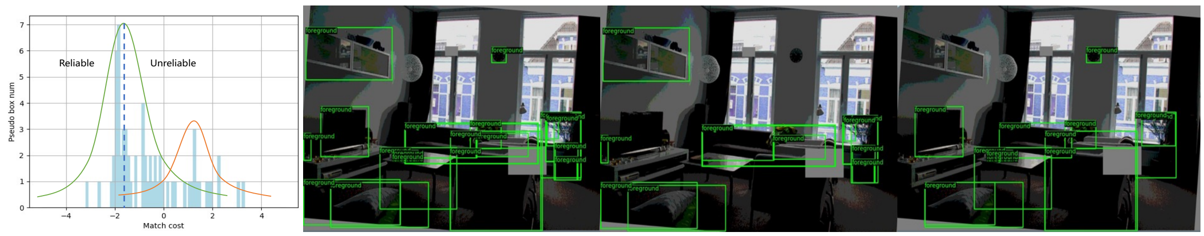
4. Experiments

4.1. Datasets and Evaluation Metrics

We validate our method on the MS-COCO benchmark [19] and Pascal VOC datasets [5]. MS-COCO contains 80 classes with 118k labeled images in the train2017 set and 123k unlabeled images in the unlabeled2017 set. In addition, the val2017 set with 5k images is provided for validation. Following [37], we consider two evaluation settings to validate our method on the MS-COCO benchmark: (1) **COCO-Partial**. 1%, 5%, and 10% images of the COCO train2017 set are randomly sampled as the labeled training data, and the remaining images of train2017 are regarded as the unlabeled data. 5 different data folds are created for each data split to validate our method. The average of standard COCO mAP on the val2017 is adopted as our final performance metric. (2) **COCO-Full**. Under this setting, the entire train2017 is utilized as the labeled data, and unlabeled2017 is used as the additional unlabeled data. The standard COCO mAP on the val2017 is taken as the evaluation metric. **Pascal VOC** contains 20 classes with VOC2007 and VOC2012 provided as the labeled data and unlabeled data respectively. The evaluation metrics are the COCO-style $AP_{50:95}$ and AP_{50} on the VOC2007 test set.

4.2. Implementation Details

To avoid loss of generality, we choose Deformable DETR [44] and DINO to integrate into our Semi-DETR



(a) Distribution of Cost (b) Initial Pseudo Boxes (c) Unreliable Pseudo Boxes (d) Reliable Pseudo Boxes

Figure 5. An illustration of Cost-based Pseudo Label Mining. We first take the image-level confidence score's mean with variance to get the initial pseudo labels, shown in (b), for each image and perform the Hungarian match to get the matching cost of each pseudo ground truth bounding box within a batch. Then, we fit a GMM model with these cost values, shown in (a). We argue that the pseudo boxes with lower cost are more likely to be the reliable pseudo boxes, so we take the lower threshold from the GMM model to filter the pseudo label again to obtain the final pseudo boxes presented in (d).

method. Following them, we use ResNet-50 [12] pre-trained on ImageNet [4] as our backbone network. Focal Loss [18] is used for classification during training. Smooth L1 Loss and GIoU [28] Loss are used for regression. We set the number of object queries to 300 for Deformable DETR and 900 for DINO, respectively. For the training hyperparameters, following [37]: (1) For the COCO-Partial benchmark, we train Semi-DETR for 120k iterations on 8 GPUs with 5 images per GPU. The first stage with one-to-many assignment is set to 60k iterations. The ratio of the labeled data and unlabeled data is set to 1:4. The weight of the unsupervised loss is set to $\alpha = 4.0$. (2) For the COCO-Full benchmark, we double the training time with COCO-unlabeled to 240k, where the first stage with one-to-many assignment is set to 180k iterations. The batch size is set to 64 on 8 GPUs with 8 images per GPU. The ratios of labeled data and unlabeled data are set to 1:1, and the loss weight of unlabeled data is set to $\alpha = 2.0$. (3) For the Pascal VOC benchmark, we train Semi-DETR for 60k iterations where The first stage with one-to-many assignment is set to 40k iterations. Other settings are kept the same with COCO-Partial. For all experiments, the confidence threshold is set to 0.4. We utilize Adam [14] with a learning rate of 0.001, and no learning rate decay scheme is used. The teacher model is updated through EMA with a momentum of 0.999. Besides, we follow the same data prepossessing, and augmentation pipeline in [37] without modifications.

4.3. Comparison with SOTA methods

We compare our Semi-DETR method with current SOTA SSOD methods on both MS-COCO and Pascal VOC datasets. We present the superiority of Semi-DETR in the following aspects: (1) comparisons to two-stage and one-stage detectors, (2) comparisons to DETR-based detectors, and (3) generalization ability.

COCO-Partial benchmark. According to Tab. 1, Semi-DETR shows significant superiority over current SOTA SSOD methods across all experiment settings in COCO-Partial. Concretely, (1) compared to SOTA two-stage and one-stage detectors, Semi-DETR outperforms PseCo (experiment 3) by 2.77, 2.00, 2.05 mAP with Deformable DETR (by 8.07, 7.60, 7.44 mAP with DINO) under the 1%, 5%, 10% settings and beats Dense Teacher (experiment 5) by 2.82, 1.49, 0.97 mAP with Deformable DETR (by 8.12, 7.09, 6.37 mAP with DINO) under the 1%, 5%, 10% settings. Obviously, Semi-DETR is a better semi-supervised object detector, and it does not require hand-crafted components used in two-stage and one-stage detectors; (2) we construct two DETR-based baselines, namely DETR under supervised training only and a simple pseudo labeling Teacher-Student architecture integrating DETR with SSOD. By comparing experiments 7-10 (or experiments 11-14), Semi-DETR outperforms the supervised baseline by 14.20, 10.80, 8.90 mAP with Deformable DETR (12.50, 10.60, 8.50 mAP with DINO) and surpasses the SSOD baseline by 5.80, 3.40, 3.30 mAP with Deformable DETR (2.10, 2.10, 1.90 mAP with DINO). This demonstrates that simply integrating DETR-based detectors with Teacher-Student architecture is not optimal. (3) we use Deformable DETR and DINO to show the generalization ability of our Semi-DETR method. Apparently, Semi-DETR consistently boosts the performance of both detectors over the corresponding baselines by clear margins (experiments 7-14). With stronger detectors like DINO, Semi-DETR still enjoys a notable performance improvement.

COCO-Full benchmark. According to Tab. 3, when adding additional unlabeled2017 data, Semi-DETR with Deformable DETR enjoys 3.6 mAP performance gain and reaches 47.2 mAP, surpassing PseCo and Dense Teacher by 1.1 and 1.1 mAP, respectively. This further manifests the effectiveness of Semi-DETR. Besides, under stronger baselines like DINO, Semi-DETR still shows obvious performance gain (+1.8 mAP), which outperforms PseCo and Dense Teacher by 4.3 and 4.3 mAP respectively and generates a new SOTA performance of 50.4 mAP.

Pascal VOC benchmark. Semi-DETR presents consistent performance improvements on the Pascal VOC benchmark as shown in Tab. 2. Generally, Semi-DETR outper-

Table 1. Comparisons with SOTA SSOD methods under the COCO-Partial setting. All results are the average of all 5 folds. Def-DETR denotes Deformable DETR. Sup Only denotes supervised only baseline.

Category	Method	ID	1%	5%	10%
Two-Stage	Unbiased Teacher	1	20.75 ± 0.12	28.27 ± 0.11	31.50 ± 0.10
	Soft-Teacher	2	20.46 ± 0.39	30.74 ± 0.08	34.04 ± 0.14
	PseCo	3	22.43 ± 0.36	32.50 ± 0.08	36.06 ± 0.24
One-Stage	DSL	4	22.03 ± 0.28	30.87 ± 0.24	36.22 ± 0.18
	Dense Teacher	5	22.38 ± 0.31	33.01 ± 0.14	37.13 ± 0.12
	Unbiased Teacher v2	6	22.71 ± 0.42	30.08 ± 0.04	32.61 ± 0.03
End-to-End	Omi-DETR(Def-DETR)	7	18.60	30.20	34.10
	Def-DETR(Sup only)	8	11.00 ± 0.24	23.70 ± 0.13	29.20 ± 0.11
	Def-DETR SSOD(Baseline)	9	19.40 ± 0.31	31.10 ± 0.21	34.80 ± 0.09
	Semi-DETR(Def-DETR)	10	**25.20 ± 0.23**	**34.50 ± 0.18**	**38.10 ± 0.14**
	DINO(Sup only)	11	18.00 ± 0.21	29.50 ± 0.16	35.00 ± 0.12
	DINO SSOD (Baseline)	12	28.40 ± 0.21	38.00 ± 0.13	41.60 ± 0.11
	Omi-DETR(DINO)	13	27.60	37.70	41.30
	Semi-DETR(DINO)	14	**30.50 ± 0.30**	**40.10 ± 0.15**	**43.50 ± 0.10**

Table 2. Comparisons with SOTA SSOD methods under the Pascal VOC setting. Def-DETR denotes Deformable DETR. Sup Only denotes supervised only baseline.

Category	Method	AP_{50}	$AP_{50:95}$
Two-Stage	Unbiased Teacher	77.37	48.69
	Soft-Teacher	-	-
	PseCo	-	-
One-Stage	DSL	80.70	56.80
	Dense Teacher	79.89	55.87
	Unbiased Teacher v2	81.29	56.87
End-to-End	Def-DETR(Sup only)	74.50	46.20
	Def-DETR SSOD(Baseline)	78.90	53.40
	Semi-DETR(Def-DETR)	**83.50**	**57.20**
	DINO(sup only)	81.20	59.60
	DINO SSOD (Baseline)	84.30	62.20
	Semi-DETR(DINO)	**86.10**	**65.20**

Table 3. Comparisons with SOTA SSOD methods under the COCO-Full setting. Def-DETR denotes Deformable DETR. Sup Only denotes supervised only baseline.

Method	100%
Unbiased Teacher	$40.2 \xrightarrow{+1.1} 41.3$
Soft-Teacher	$40.9 \xrightarrow{+3.6} 44.5$
PseCo	$41.0 \xrightarrow{+5.1} 46.1$
DSL	$40.2 \xrightarrow{+3.6} 43.8$
Dense Teacher	$41.2 \xrightarrow{+3.6} 46.1$
Semi-DETR(Def-DETR)	$\mathbf{43.6} \xrightarrow{+3.6} \mathbf{47.2}$
Semi-DETR(DINO)	$\mathbf{48.6} \xrightarrow{+1.8} \mathbf{50.4}$

forms the supervised baseline by 9.0 on AP_{50} and 11.0 on $AP_{50:95}$ with Deformable DETR (by 4.9 on AP_{50} and 5.6 on $AP_{50:95}$ with DINO). Furthermore, Semi-DETR beats all previous SOTA SSOD methods by significant margins on both AP_{50} and $AP_{50:95}$.

4.4. Ablation Study

We conduct extensive experiments to verify the effectiveness of Semi-DETR in the following aspects: (1) component effectiveness; (2) variants of Stage-wise Hybrid Matching (SHM); (3) effectiveness of Cross-view Query Consistency (CQC) and Cost-based Pseudo Label Mining (CPM); (4) hyper-parameters. All experiments are performed with DINO as the base detector on the 10% labeled images setting of the COCO-Partial benchmark.

Component Effectiveness. According to Tab. 4, we perform four experiments to verify the effectiveness of each proposed component. We formulate a strong baseline that integrates DINO with SSOD via pseudo labeling in experiment 1. In general, our proposed components enjoy consistent performance improvements. Specifically, by introducing the SHM module, it outperforms the baseline by 1.1 mAP. Further integrating the CQC and CPM modules brings an extra 0.8 improvement. This shows that our proposed components are complementary to each other and proves the effectiveness of each component in our model.

Table 4. Component effectiveness of Semi-DETR. SHM denotes the Stage-wise Hybrid Matching, CQC means Cross-view Query Consistency, and CPM represents Cost-based Pseudo Label Mining, respectively.

ID	SHM	CQC	CPM	mAP	AP_{50}	AP_{75}
1				41.6	58.3	45.1
2	✓			42.7	59.3	46.2
3	✓	✓		43.1	59.6	46.6
4	✓	✓	✓	**43.5**	**59.7**	**46.8**

Variants of SHM. We examine the impact of different one-to-many assignment strategies within SHM in the first stage of training. Concretely, Max-IoU [27], ATSS [41] and SimOTA [8] are chosen as the alternatives. All models are trained for 60k iterations. As presented in Tab. 6, it is interesting to find that not all traditional one-to-many assignment methods are effective in DETR-based detectors. Max-IoU assignment strategy and ATSS show significant performance degradation when applied to the first stage, even though they are commonly used in traditional object detectors. On the other hand, SimOTA shows comparable performance to our one-to-many assignment strategy. This is possibly caused by the fact that SimOTA and our method adopt a ranking-based one-to-many assignment strategy while Max-IoU and ATSS utilize hard or dynamic thresholding-based one-to-many assignment strategy, which leads to a significant difference number of assigned positive samples for each pseudo ground truth bounding box and thus suffers performance degradation. More analysis can be found in the supplementary document.

Effectiveness of CQC+CPM. According to Tab. 5, we compare four different methods to generate pseudo labels for CQC and evaluate the precision and recall metrics of the generated pseudo labels. First, we present two methods (by setting a fixed classification score $\tau_s = 0.4$ or by selecting Top-K pseudo labels with the highest confidence scores) that obtain pseudo labels with high precision (81.5% or 80.2%) and low recall (41.3% or 39.4%) but observe marginal performance gains. Then we present the Mean+Std method that aims to balance the precision (60.2%) and recall (54.0%) of pseudo labels via combining the image-level mean confidence score and variance $\tau = \mu + \sigma$, which enjoys a better performance improvement (+0.4 mAP). Finally, our Cost-based GMM method achieves a better trade-off between the precision (77.6%) and recall (52.1%) metrics, which has a 0.8 performance gain.

Table 5. Effects of different methods to filter pseudo labels for cross-view consistency training.

Method	mAP	Precision	Recall
Fixed(0.4)	42.8	81.5%	41.3%
Top-K(K=9)	42.9	80.2%	39.4%
Mean + Std	43.1	60.2%	54.0%
Cost-based GMM	43.5	77.6%	52.1%

Table 6. Effects of the different one-to-many assignment methods in the first stage.

Strategy	mAP	AP_{50}	AP_{75}
Max-IoU	11.4	15.0	12.1
ATSS	18.7	30.5	18.9
SimOTA	42.5	**59.9**	45.2
Ours	**42.8**	59.8	**46.0**

Hyperparameters. We study two types of hyperparameters in our model: (1) the pseudo label threshold τ_s; (2) the training iterations T_1 of the first stage in SHM. For τ_s, according to Tab. 8, the best performance is achieved when $\tau_s = 0.4$. Possibly, a lower threshold could introduce noisy pseudo labels, while a higher threshold could decrease the effective number of pseudo labels. For the training iterations of the first stage, according to Tab. 7, performing the one-to-many assignment strategy across both stages achieves 44.0 mAP at the cost of using NMS in the end. The appropriate training time of the first stage is at 60k iterations, which achieves the best performance of 43.5 mAP and does not require NMS post-process at the same time.

Table 7. Effects of the training iteration T_1 of the first stage using one-to-many assignment strategy in Stage-wise Hybrid Matching.

T_1	40k	60k	80k	100k	120k
mAP	42.9	**43.5**	43.2	43.0	**44.0**
NMS-Free	Y	Y	Y	Y	N

Table 8. Effects of the pseudo label threshold τ_s.

τ_s	0.2	0.3	0.4	0.5	0.6
mAP	42.6	43.0	**43.5**	43.2	42.8

5. Conclusion

We analyzed the challenges of the DETR-based object detectors on semi-supervised object detection, including the learning inefficiency of one-to-one assignment with inaccurate pseudo labels and the difficulties of designing consistency-based regularization due to the absence of deterministic correspondence from object queries. We proposed Semi-DETR, the first transformer-based end-to-end semi-supervised object detector. It consists of a Stage-wise Hybrid Matching method that embraces the merits of both one-to-many assignment and one-to-one assignment strategies, a Cross-view Query Consistency method that learns semantic feature invariance of object queries from different views via unlabeled images, and a Cost-based Pseudo Labeling module that adaptively mines more reliable pseudo labels for improving the efficiency of consistency training. Extensive experiments demonstrate the superiority of Semi-DETR on both MS-COCO and Pascal VOC benchmarks.

Acknowledgments

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation (NO. 2020B1515020048), in part by the National Natural Science Foundation of China (NO. 61976250), in part by the Shenzhen Science and Technology Program (NO. JCYJ20220530141211024) and in part by the Fundamental Research Funds for the Central Universities under Grant 22lgqb25.

References

[1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[3] Binghui Chen, Pengyu Li, Xiang Chen, Biao Wang, Lei Zhang, and Xian-Sheng Hua. Dense learning based semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4815–4824, 2022.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[5] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, Jan. 2015.

[6] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3490–3499. IEEE Computer Society, 2021.

[7] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 303–312, 2021.

[8] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[9] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[10] Qiushan Guo, Yao Mu, Jianyu Chen, Tianqi Wang, Yizhou Yu, and Ping Luo. Scale-equivalent distillation for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14522–14531, 2022.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *Advances in neural information processing systems*, 32, 2019.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022.

[16] Gang Li, Xiang Li, Yujie Wang, Shanshan Zhang, Yichao Wu, and Ding Liang. Pseco: Pseudo labeling and consistency training for semi-supervised object detection. *arXiv preprint arXiv:2203.16317*, 2022.

[17] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33:21002–21012, 2020.

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[20] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022.

[21] Yen-Cheng Liu and et al. Unbiased teacher for semi-supervised object detection. In *ICLR 2021*.

[22] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9819–9828, 2022.

[23] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3651–3660, 2021.

[24] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

[25] Yuxiang Nie, Chaowei Fang, Lechao Cheng, Liang Lin, and Guanbin Li. Adapting object size variance and class imbalance for semi-supervised object detection. In *Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[26] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

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

[28] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding

box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[29] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020.

[30] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[31] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[33] Kuo Wang, Yuxiang Nie, Chaowei Fang, Chengzhi Han, Xuewen Wu, Xiaohui Wang, Liang Lin, Fan Zhou, and Guanbin Li. Double-check soft teacher for semi-supervised object detection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.

[34] Kuo Wang, Jingyu Zhuang, Guanbin Li, Chaowei Fang, Lechao Cheng, Liang Lin, and Fan Zhou. De-biased teacher: Rethinking iou matching for semi-supervised object detection. In *Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[35] Pei Wang and et al. Omni-detr: Omni-supervised object detection with transformers. In *CVPR 2022*.

[36] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[37] Mengde Xu and et al. End-to-end semi-supervised object detection with soft teacher. In *ICCV 2021*.

[38] Qize Yang, Xihan Wei, Biao Wang, Xian-Sheng Hua, and Lei Zhang. Interactive self-training with mean teachers for semi-supervised object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5941–5950, 2021.

[39] Zhuyu Yao, Jiangbo Ai, Boxun Li, and Chi Zhang. Efficient detr: improving end-to-end object detector with dense prior. *arXiv preprint arXiv:2104.01318*, 2021.

[40] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[41] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9759–9768, 2020.

[42] Hongyu Zhou, Zheng Ge, Songtao Liu, Weixin Mao, Zeming Li, Haiyan Yu, and Jian Sun. Dense teacher: Dense pseudo-labels for semi-supervised object detection. *arXiv preprint arXiv:2207.02541*, 2022.

[43] Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4081–4090, 2021.

[44] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.