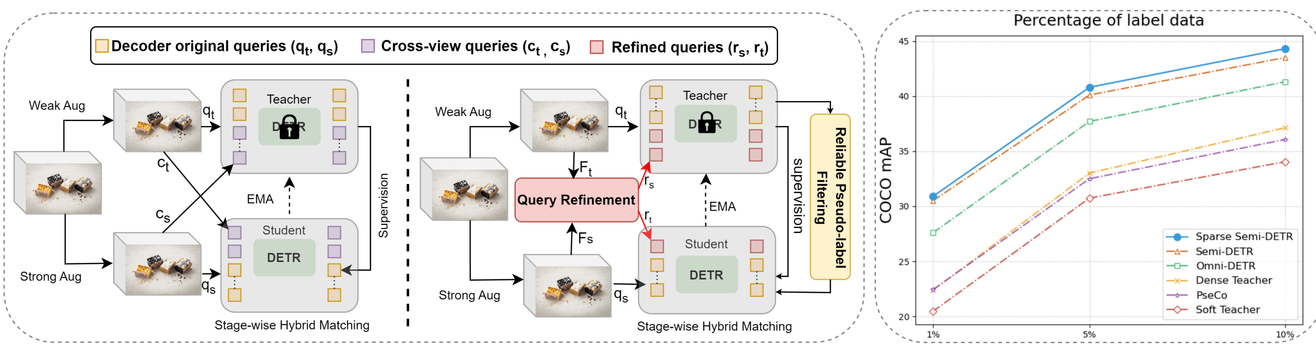
Sparse Semi-DETR: Sparse Learnable Queries for Semi-Supervised Object Detection

Tahira Shehzadi[1,2,3], Khurram Azeem Hashmi[1,2,3], Didier Stricker[1,2,3], Muhammad Zeshan Afzal[1,2,3]
[1]DFKI, [2]RPTU Kaiserslautern-Landau, [3]MindGarage-RPTU



(a) Semi-DETR (b) Sparse Semi-DETR (c) Comparison of SSOD Methods

Figure 1. (a)-(b) Comparative Overview of SSOD advancements: Sparse Semi-DETR's major improvement lies in its Query Refinement Module and Reliable Pseudo-label Filtering Module, significantly enhancing detection of small or obscured objects and reliability in complex scenarios, surpassing all other methods as shown in the graph in **(c)**.

Abstract

In this paper, we address the limitations of the DETR-based semi-supervised object detection (SSOD) framework, particularly focusing on the challenges posed by the quality of object queries. In DETR-based SSOD, the one-to-one assignment strategy provides inaccurate pseudo-labels, while the one-to-many assignments strategy leads to overlapping predictions. These issues compromise training efficiency and degrade model performance, especially in detecting small or occluded objects. We introduce Sparse Semi-DETR, a novel transformer-based, end-to-end semi-supervised object detection solution to overcome these challenges. Sparse Semi-DETR incorporates a ***Query Refinement Module*** *to enhance the quality of object queries, significantly improving detection capabilities for small and partially obscured objects. Additionally, we integrate a* ***Reliable Pseudo-Label Filtering Module*** *that selectively filters high-quality pseudo-labels, thereby enhancing detection accuracy and consistency. On the MS-COCO and Pascal VOC object detection benchmarks, Sparse Semi-DETR achieves a significant improvement over current state-of-the-art methods that highlight Sparse Semi-DETR's effectiveness in semi-supervised object detection, particularly in challenging scenarios involving small or partially obscured objects.*

1. Introduction

Semi-Supervised Object Detection (SSOD) aims to improve the effectiveness of fully supervised object detection through the integration of abundant unlabeled data [3, 12, 15, 19, 28, 39, 49–52, 58]. It has applications in diverse fields, ranging from autonomous vehicles [14, 20] to healthcare [31, 44], where obtaining extensive labeled datasets is often impractical or cost-prohibitive [1].

Several SSOD methods [3, 12, 15, 19, 28, 39, 49–52, 58] have been proposed. Two prevalent approaches in this domain are pseudo-labeling [27, 30, 39, 45, 50–52, 58] and consistency-based regularization [3, 12, 15, 19, 28, 49]. STAC [39] introduced a simple multi-stage SSOD training method with pseudo-labeling and consistency training, later simplified by a Teacher-Student framework for generating pseudo-labels [27]. Based on this framework, considerable research efforts have been directed towards enhancing the quality of pseudo-labels [51, 58]. These traditional SSOD methods are built upon conventional detectors like one-stage [33, 42] and two-stage [9, 35], which involve various manually designed components such as anchor boxes and non-maximum suppression (NMS). Employing object detection methods in SSOD poses several potential challenges that must be carefully dealt with to obtain reasonable performance. These factors include overfitting of the labeled data [37], pseudo-label noise [10], bias induced through label imbalance [17, 32], and poor detection performance on small objects [56]. Recently, DETR-based [2, 16, 18, 25, 38, 55, 59] SSOD methods [46, 56] remove the need for traditional components like NMS.

Even though DETR-based SSOD [46, 56] has progressed remarkably, state-of-the-art methods possess some limitations. (1) DETR-based SSOD methods perform poorly in the detection of small objects, as shown in Figure 4. This is because these methods don't use multi-scale features [36] like Feature Pyramid Networks (FPN) [22], which play an important role in identifying smaller objects as in CNN-based SSOD methods [3, 12, 15, 19, 28, 49, 51, 52, 58]. Although recent advancements in DETR-based object detection [2, 16, 18, 25, 38, 55, 59] have improved the detection of small objects, their SSOD adaptation is still unable to cater this challenge effectively [56]. (2) SSOD approaches [51, 52, 56, 58] rely on handcrafted post-processing methods such as NMS [35]. This problem specifically appears in DETR-based SSOD when we use a large number of object queries and the one-to-many assignment strategy [56]. In DETR-based SSOD methods, this problem is partially solved using the one-to-one or hybrid (combination of one-to-one and one-to-many) assignment strategy. However, the hybrid assignment strategy is preferred because the one-to-one assignment strategy produces inaccurate pseudo-labels [46], thus resulting in inefficient learning. Although the number of duplicate bounding boxes is less in the hybrid strategy [56], the amount is high enough to impact object detection performance adversely, as depicted in Figure 5. (3) The pseudo-label generation produces both high and low-quality labels. The DETR-based SSOD methods lack an effective refinement strategy for one-to-many assignments, which is crucial for filtering out low-quality proposals.

To address the above mentioned issues, we propose enhancing the state-of-the-art DETR-based SSOD approach, namely 'Sparse Semi-DETR', presented in Figure 1 (b). Our approach involves expanding its architecture by integrating a couple of novel modules designed to mitigate the identified shortcomings. The key module among these is the **Query-Refinement module**, as depicted in Figure 2 and explained in Figure 3. This module significantly improves the quality of the queries and reduces their numbers. The proposed module uses the low-level features from the backbone and high-level features extracted directly from weakly augmented images using ROI alignment [13]. Fusing these features results in overcoming the first shortcoming, i.e., detecting small and obscured objects, as shown in Figure 4. The attention mechanism drives the aggregation of the features, resulting in refined, high-quality features to carry forward. To ensure the quality of the query features, the attention mechanism is accompanied by a query-matching strategy for filtering irrelevant queries. Thus, the Query Refinement Module not only improves the quality of the queries but also reduces their numbers, giving rise to efficient processing. This module results in significantly fewer overlapping proposals, improving the performance overall, thereby solving the second limitation. Besides, we introduce a **Reliable Pseudo-Label Filtering Module**, as illustrated in Figure 2, inspired by Hybrid-DETR [16] to address the third limitation. Employing this module significantly reduces the low-quality pseudo-labels. Therefore, it further reduces the amount of duplicate predictions that may still occur after the second stage of the hybrid assignment strategy. Our approach provides better results than previous SSOD methods, as shown in Figure 1 (c).

The key contributions of this work can be outlined as follows:

1. We present Sparse Semi-DETR, a novel approach in semi-supervised object detection, introducing two novel contributions. To our knowledge, we are the first to examine and propose query refinement and low-quality proposal filtering for the one-to-many query assignment strategy.
2. We introduce a novel query refinement module designed to improve object query features, particularly in complex detection scenarios such as identifying small or partially obscured objects. This enhancement not only boosts performance but also aids in learning semantic feature invariance among object queries.
3. We introduce a Reliable Pseudo-Label Filtering Module specifically designed to reduce the effect of noisy pseudo-labels. This module is designed to efficiently identify and extract reliable pseudo boxes from unlabeled data using augmented ground truths, enhancing the consistency of the learning process.
4. Sparse Semi-DETR outperforms current state-of-the-art methods on MS-COCO and Pascal VOC benchmarks. With only 10% labeled data from MS-COCO using ResNet-50 backbone, it achieves a 44.3 mAP, exceeding prior baselines by 0.8 mAP. Additionally, when trained on the complete COCO set with extra unlabeled data, it further improves, rising from 49.2 to 51.3 mAP.

2. Related Work

2.1. Object Detection

Object detection identifies and locates objects in images or videos. Deep learning-based object detection approaches are typically categorized into two primary groups: two-stage detectors [9, 35] and one-stage detectors [23, 26, 34, 42]. These methods depend on numerous heuristics, such as generating anchors and NMS. Recently, DEtection TRansformer (DETR) [2] considers object detection as a set prediction problem, using transformer [43] to adeptly transform sparse object candidates [40] into precise target objects. Our Sparse Semi-DETR detects small or partially obscured objects in the DETR-based SSOD setting. Notably, our framework is compatible with various DETR-based detectors [4, 5, 7, 8, 24, 29, 47, 48, 59], offering flexibility in integration.

2.2. Semi-Supervised Object Detection

Most research in SSOD employs detectors categorized into three types: one-stage, two-stage, and DETR-based systems. **One-stage** STAC [39], an early SSOD, introduced a simple training strategy combining pseudo-labeling and consistency training, later streamlined by a student-teacher framework for easier pseudo-label generation [27]. DSL [3] introduced novel techniques including Adaptive Filtering, Aggregated Teacher, and uncertainty-consistency-regularization for improved generalization. Dense Teacher [57] introduced Dense Pseudo-Labels (DPL) for richer information and a region selection method to reduce noise.

Two-stage. Humble Teacher [41] uses soft labels and a teacher ensemble to boost pseudo-label reliability, matching other results. Instant-Teaching [58] creates pseudo annotations from weak augmentations, treating them as ground truth under strong augmentations with Mixup [54]. Unbiased Teacher [27] tackles class imbalance in pseudo-labeling with focal loss, focusing on underrepresented classes. Soft Teacher [51] minimizes incorrect foreground proposal classification by applying teacher-provided confidence scores to reduce classification loss. PseCo [19] enhances detector performance by combining pseudo-labeling with label and feature consistency methods, also using focal loss to address class imbalance.

DETR-based. Omni-DETR [46] is designed for omni-supervised detection and adapts to SSOD with a basic pseudo-label filtering method. It employs the one-to-one assignment strategy proposed in DETR [2], and encounters challenges when dealing with inaccurate pseudo-bounding boxes produced by the teacher network. These inaccuracies result in reduced performance, highlighting its limitations. Semi-DETR [56] adopts a stage-wise strategy, employing a one-to-many matching strategy in the first stage and switching to a one-to-one matching strategy in the second stage. This approach provides NMS-free end-to-end detection benefits but reduces performance compared to a one-to-many assignment strategy. Moreover, omni-DETR and Semi-DETR struggle to detect small or occluded objects. Our work introduces an advanced query refinement module that significantly refines object queries, enhancing training efficiency and performance and leading to the detection of small or densely packed objects in the DETR-based SSOD framework.

3. Preliminary

In DETR-based SSOD, one-to-one assignment strategy, denoted by $\hat{\sigma}_{one2one}$, is achieved by applying Hungarian algorithm between the predictions made by the student model and the pseudo-labels provided by the teacher model as follows:

$$\hat{\sigma}_{one2one} = \underset{\sigma \in \xi_N}{\arg\min} \sum_{j}^{N} \mathcal{L}_{\text{match}} \left(\hat{y}_j^t, \hat{y}_{\sigma(j)}^s \right) \tag{1}$$

where $\mathcal{L}_{\text{match}} \left(\hat{y}_j^t, \hat{y}_{\sigma(j)}^s \right)$ is the matching cost between the pseudo-labels $\hat{y}_j^t$ generated by the teacher network and the predictions of the student network with index $\sigma(j)$ and ξ_N is the permutation of N elements. Semi-DETR [56] addresses the issue of imprecise initial pseudo-labels by shifting from a one-to-one to a one-to-many assignment strategy, increasing the number of positive object queries to improve detection accuracy:

$$\hat{\sigma}_{one2many} = \left\{ \underset{\sigma_j \in C_M^N}{\arg\min} \sum_{k}^{M} \mathcal{L}_{\text{match}} \left(\hat{y}_j^t, \hat{y}_{\sigma_j(k)}^s \right) \right\}_j^{|\hat{y}^t|} \tag{2}$$

where C_N^M represents the combination of M and N, denoting that a subset of M proposals is associated with each pseudo box $\hat{y}_j^t$. Semi-DETR initially adopts a one-to-many assignment to improve label quality, then shifts to one-to-one assignment for an NMS-free model. This approach adopts a one-to-many assignment strategy aimed at boosting performance, but it's less effective with small or occluded objects.

4. Sparse Semi-DETR

In semi-supervised learning, a collection of labeled data denoted as D_l, where $D_l = \{x_i^l, y_i^l\}_{i=1}^{N_l}$ is given, along with a set of unlabeled data represented as D_u, where $D_u = \{x_i^u\}_{i=1}^{N_u}$. Here, N_l and N_u correspond to the number of labeled and unlabeled data. The annotations y_i^l for the label data x^l contain object labels and bounding box information. The pipeline of the Sparse Semi-DETR framework is depicted in Figure 2. It introduces a Query Refinement Module for processing query features to enhance semantic representation for complex detection scenarios, such as identifying small or partially obscured objects. Additionally, we integrate a Reliable Pseudo-Label Filtering Module that selectively filters high-quality pseudo-labels, thereby enhancing detection accuracy. For comparison purposes, we employ DINO [55] with a ResNet-50 backbone. This section gives a detailed overview of the modules of Sparse Semi-DETR. We explain briefly our semi-supervised approach in Appendix A1.1.

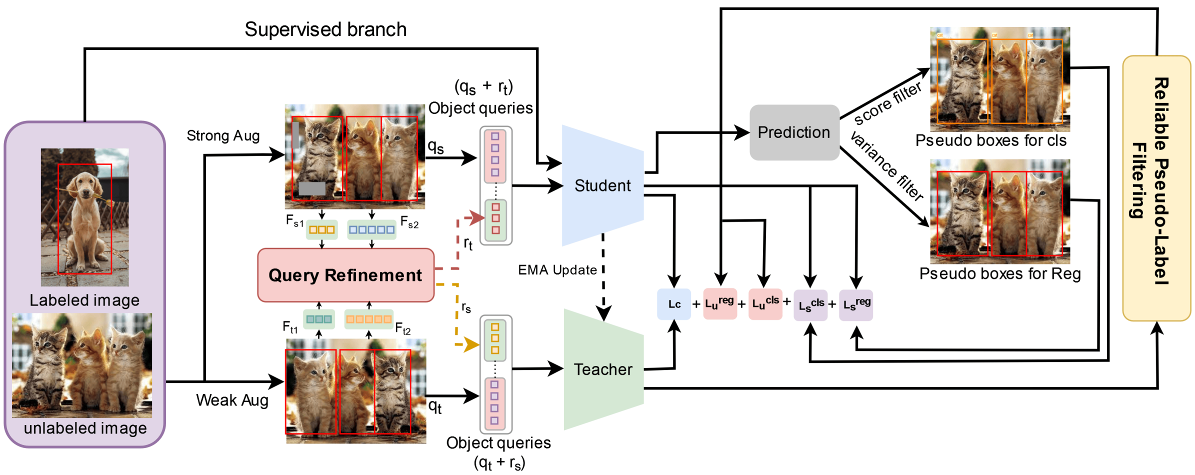


Figure 2. An overview of the Sparse Semi-DETR framework. It contains two networks: the student network and the teacher network. Labeled data is used for student network training, employing a supervised loss. Unlabeled data is fed to the teacher network with weak augmentation and the student network with strong augmentation. The teacher network takes unlabeled data to generate pseudo-labels. Here, the query refinement module provides refined queries to avoid incorrect bipartite matching with teacher-generated pseudo-labels. For a detailed overview of the query refinement module, see Figure 3. Furthermore, a Reliable Pseudo-label Filtering strategy is employed to filter low-quality pseudo-labels progressively during training.

4.1. Query Refinement

Inspired by recent advancements in vision-based networks [11, 53], we introduce an innovative approach to enhance object query features. For each unlabeled image $I \in \mathbb{R}^{H\times W\times C}$, we extract query features $F_{s1} \in \mathbb{R}^{b\times W_1\times 256}$ from strongly augmented image I. Similarly, we extract query features F_{t1} from weakly augmented image I, also in the same dimension. Subsequently, feature extraction from the image backbones occurs. This results in the generation of features F_{s2} for the student and F_{t2} for the teacher network as $\mathbb{R}^{b\times W_2\times 256}$. These features, encompassing both label and bounding box details, vary with the batch size, as indicated by b. The feature sets W_1 and W_2 differ in size, with W_2 being substantially larger than W_1. We provide a brief overview of each component of Query Refinement Module as illustrated in Figure 3.

Query Refinement Module. In our approach, we handle multi-scale features F_{t1} and F_{t2} with a focus on effective aggregation. The finer details are encapsulated within the features F_{t1}, while the features F_{t2} encapsulate more abstract elements such as shapes and patterns. Simple aggregation of these features has been shown to degrade performance, as indicated in Table 5e). To solve this issue, we implement dual strategies to extract local and global information from high and low-resolution features. High-resolution features are crucial for detecting small objects. However, processing them with attentional operations is computationally demanding. To address this, we firstly convert the query label features $F_{t2} \in (\mathbb{R}^{b\times W_2\times 256})$ into $F'_{t2} \in (\mathbb{R}^{b\times W_2\times 16})$ by decreasing the channel dimension, and retaining the original resolution $b \times W_2$. Then, we apply attentional mechanism on F'_{t2} to calculate the attentional weights W_{k+q} in attention block as follows:

$$W_{k+q} = F'_k \cdot F'_q, \tag{3}$$

$$\bar{\mathbf{W}}_{k+q} = \frac{\exp(W_{k+q})}{\sum_{l=1}^{L}\exp(W_{k+q})}, \tag{4}$$

where W_{k+q} is the attentional weights of F'_k and F'_q, and $\bar{\mathbf{W}}_{k+q}$ is the normalized form of W_{k+q}. Using normalized attention weights, we compute the enhanced queries representation Q as follows:

$$Q = \bar{\mathbf{W}}_{k+q} \cdot F'_v \tag{5}$$

now we find the similarity between the attentional F_{t2} features and F_{t1} features to obtain $F'_{cs} \in \mathbb{R}^{b\times W_1\times 16}$ from $Q \in \mathbb{R}^{b\times W_2\times 16}$ as follows:

$$F'_{cs} = \frac{\sum_{i=l}^{n} P_l Q_l}{\sqrt{\sum_{l=1}^{n} P_l^2}\sqrt{\sum_{l=1}^{n} Q_l^2}} \tag{6}$$

where P and Q are F_{t1} and attentional F_{t2} features, respectively. Then, we concatenate F'_{cs} with P to obtain refined query features. Interestingly, we observe a performance drop when our feature refinement strategy is applied to strongly augmented image features for the teacher network, as detailed in Table 5b. However, we achieve optimal results by concatenating strongly augmented image features and applying our refinement strategy to weakly augmented image features. Consequently, we proceed by concatenating the features F_{s1} with F_{s2}, thereby obtaining the query features

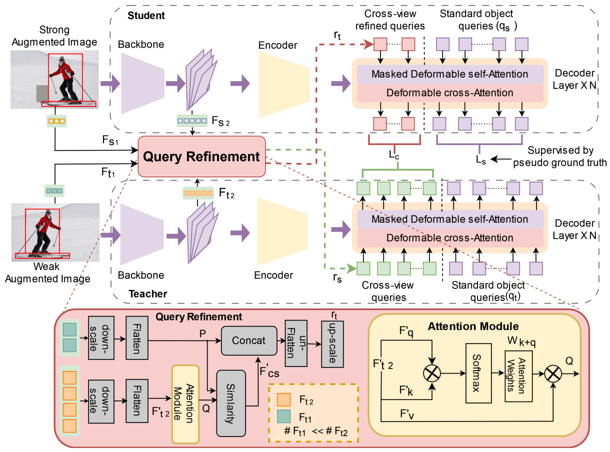


Figure 3. Overview of the Query Refinement Module. The query features from strong and weak augmented unlabeled images are refined through the Query Refinement Module. It amplifies the semantic representation of object queries and improves performance for small objects. For the best view, zoom in.

r_s. Note that r_t, despite having a dimensional size equivalent to $F'_{cs} + P$, encapsulates substantially more intricate representations. This improved performance is due to the integration of high-resolution and low-resolution features.

Then, we form the decoder queries in the student-teacher network by merging the teacher's original queries q_t with refined queries r_s, and the student's original queries q_s with refined queries r_t, respectively. This integration forms the inputs for the decoder as follows:

$$\hat{o}_t, o_t = \text{Dec}_t([r_s, q_t], E_t[A]) \tag{7}$$

$$\hat{o}_s, o_s = \text{Dec}_s([r_t, q_s], E_s[A]) \tag{8}$$

Where E_s and E_t refer to the encoded image features. $\hat{o}_s$ and $\hat{o}_t$ indicate the decoded features of refined queries, while o_s and o_t represent the decoded features of original object queries. Here, t is for teacher and s for student network. Following DN-DETR [18], we use the attention mask A to protect information leakage, ensuring the learning process's integrity.

4.2. Reliable Pseudo-Label Filtering Module

The one-to-many training strategy, while effective, causes duplication prediction in the first stage. We introduce a pseudo-label filtering module to address this and improve the filtering of pseudo boxes rich in semantic content for refined query learning. This module is designed to efficiently identify and extract reliable pseudo boxes from unlabeled data using augmented ground truths. We employ m groups of ground truths $\hat{g} = \{\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_m\}$ for one-to-many assignment strategy and select the top-k predictions as follows:

$$\hat{\sigma}_{one2many} = \left\{ \underset{\sigma_j \in C_M^N}{\arg\min} \sum_{k}^{M} \mathcal{L}_{\text{match}} \left(\hat{y}_j^t, \hat{g}_{\sigma_j(k)}^s \right) \right\}_j^{|\hat{y}^t|} \tag{9}$$

where C_N^M represents the combination of M and N, denoting that a subset of M proposals is associated with each pseudo box $\hat{y}_j^t$. Here, m is set to 6. Furthermore, we use the remaining predictions to filter out duplicates in the top-k predictions in the one-to-one matching branch. Through this improved selection scheme, we achieve a performance improvement of 0.4 mAP when m is set to 6, as shown in Table 5a. However, we observe no significant benefits when increasing m greater than 6, as detailed in Table 6a.

5. Experiments

5.1. Datasets

We evaluate our approach on the MS-COCO [21] and Pascal VOC [6] datasets, benchmarking it against current SOTA SSOD methods. Following [51, 56], Sparse Semi-DETR is evaluated in three scenarios: **COCO-Partial.** We use 1%, 5%, 10% of `train2017` as label data and rest as unlabeled data. **COCO-Full.** We take `train2017` as label data and `unlabel2017` as unlabel data. **VOC.** We take `VOC2007` as label data and `VOC2012` as unlabel data. Evaluation metrics include $AP_{50:95}$, AP_{50}, and AP_{75} [51, 56].

5.2. Implementation Details

We set the quantity of DINO original object queries to 900. For the setting hyperparameters, following [56] : (1) In the COCO-Partial setting, we set the training iterations to 120k with a labeled to unlabeled data ratio of 1:4. The first 60k iterations adopt a one-to-many assignment strategy. (2) In the COCO-Full setting, Sparse Semi-DETR is trained for 240k iterations with labeled to unlabeled data ratio of 1:1. The first 120k iterations adopt a one-to-many assignment strategy. (3) In the VOC setting, we train the network for 60k iterations with a labeled to unlabeled data ratio of 1:4. The first 40k iterations adopt a one-to-many assignment strategy. For all experiments, the filtering threshold σ value is 0.4. We set the value of m to 6 and the value of k to 4. We provide complete implementation details for each experiment in Appendix A1.2.

6. Results and Comparisons

We evaluate Sparse Semi-DETR and compare it against current SOTA SSOD methods. Our results demonstrate the superior performance of Sparse Semi-DETR in these aspects: (1) its effectiveness compared to both one-stage and two-stage detectors, (2) its comparison with traditional DETR-based detectors, and (3) its exceptional proficiency in accurately

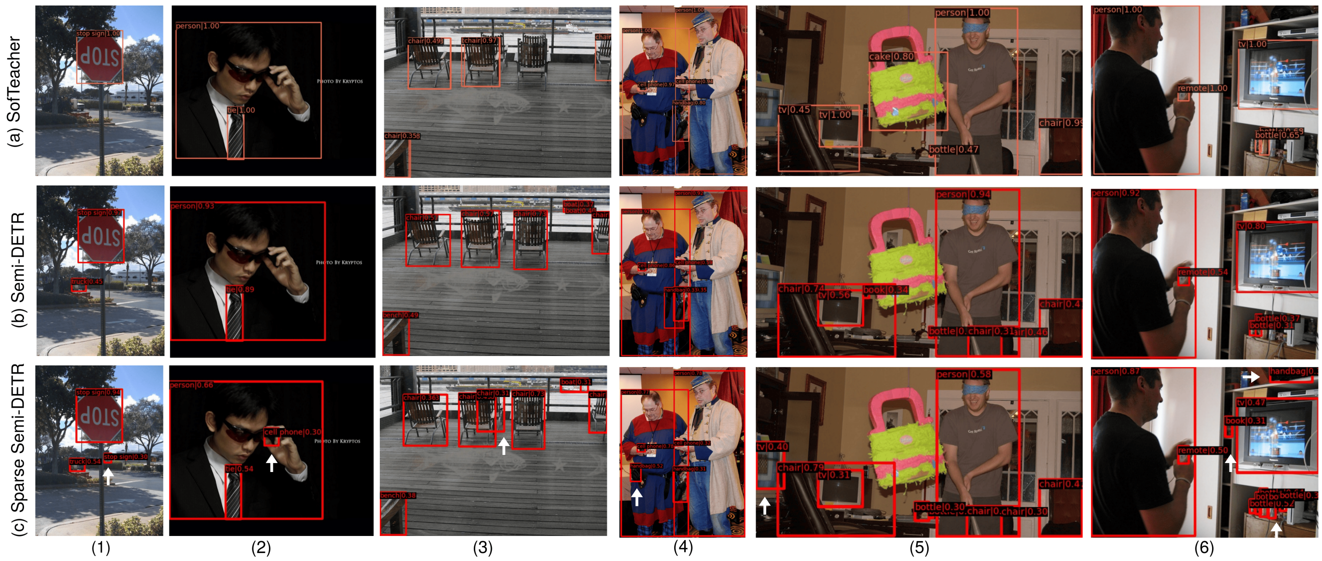


Figure 4. Visual comparison of Sparse Semi-DETR with the two previous approaches on the COCO 10% label dataset. These results highlight Sparse Semi-DETR's capabilities, particularly in identifying small objects and those obscured by obstacles (as indicated by white arrows) in the third-row images. For optimal clarity and detail, please zoom in.

detecting small and partially occluded objects. We provide more results details in Appendix A1.3.

COCO-Partial benchmark. Sparse Semi-DETR outperforms the current SSOD methods in COCO-Partial across all experiment settings, as demonstrated in Table 1. (1) We compare our method to both one-stage and two-stage SSOD. Sparse Semi-DETR surpasses Dense Teacher by 8.52, 7.79, 7.17 mAP on 1%, 5%, and 10% label data. It also outperforms PseCo by 8.47, 8.30, 8.24 mAP on 1%, 5%, and 10% label data. Sparse Semi-DETR's superior performance as a semi-supervised object detector is achieved without needing hand-crafted components commonly used in two-stage and one-stage detectors. (2) When compared to DETR-based detectors, Sparse Semi-DETR outperforms omni-DETR by 3.30, 3.10, and 3.00 mAP and beats Semi-DETR by 0.40, 0.70, 0.80 mAP on 1%, 5%, and 10% label data. (3) Sparse Semi-DETR's exceptional proficiency in precisely detecting small and partially obscured objects is a standout feature. In Figure 4, we visually compare Sparse Semi-DETR with the two preceding approaches using the COCO 10% labeled dataset. These results demonstrate the impressive capabilities of Sparse Semi-DETR, particularly in its ability to identify small objects and objects concealed by obstacles, as highlighted in the third-row images by the white arrows. Table 2 exhibits a remarkable performance boost of Sparse Semi-DETR on small objects. It surpasses the Semi-DETR by 1.20, 0.90 and 1.70 mAP on small objects using 1%, 5%, and 10% label data, respectively. It highlights the superior efficiency and accuracy of Sparse Semi-DETR in detecting smaller objects.

COCO-Full benchmark. In Table 4, when incorporating additional `unlabel2017` data, Sparse Semi-DETR demonstrates a substantial improvement, achieving an impressive 2.1 mAP gain and reaching a total of 51.3 mAP. It surpasses the performance of Dense Teacher, PseCo, and Semi-DETR by 5.2, 5.2, and 0.9 mAP, respectively, highlighting the effectiveness of Sparse Semi-DETR.

Pascal VOC benchmark. Sparse Semi-DETR exhibits a remarkable performance boost on the Pascal VOC benchmark, as shown in Table 3. It surpasses the supervised baseline by 5.1 improvement on AP_{50} and by 5.91 impressive increase on $AP_{50:95}$. Furthermore, it outperforms all previously single-stage, two-stage, and DETR-based SSOD methods by a significant margin.

6.1. Ablation Studies

This section ablates the key design choices of Sparse Semi-DETR. The experiments detailed in this section are executed on the MS COCO dataset with 10% label data, employing DINO as the primary detector.

Effect of Individual Component. We conduct three experiments to assess the efficacy of each module of Sparse Semi-DETR, as detailed in Table 5a. We employ Semi-DETR as our baseline model. Notably, integrating each component into Sparse Semi-DETR leads to consistent performance improvements. Specifically, adding the Query Refinement and Pseudo-Label Filtering modules yields a significant increase of 0.8 mAP. These results confirm that each component of Sparse Semi-DETR enhances our model's performance.

Effect of Query Refinement Module. We examine the impact of the Query Refinement (QR) Module. In Table 5b, we explore various QR combinations to determine the most

Methods	Reference	COCO-Partial		
		1%	5%	10%
FCOS [42] (Supervised)	-	8.43 ± 0.03	17.01 ± 0.01	20.98 ± 0.01
DSL [3]	CVPR22	22.03 ± 0.28 (+13.98)	30.87 ± 0.24 (+13.86)	36.22 ± 0.18 (+15.24)
Unbiased Teacher v2 [28]	CVPR22	22.71 ± 0.42 (+14.28)	30.08 ± 0.04 (+13.07)	32.61 ± 0.03 (+11.63)
Dense Teacher [57]	ECCV22	22.38 ± 0.31 (+13.95)	33.01 ± 0.14 (+16.00)	37.13 ± 0.12 (+16.15)
Faster RCNN [35] (Supervised)	-	9.05 ± 0.16	18.47 ± 0.22	23.86 ± 0.81
Humble Teacher [41]	CVPR22	16.96 ± 0.38 (+7.91)	27.70 ± 0.15 (+9.23)	31.61 ± 0.28 (+7.75)
Instant-Teaching [58]	CVPR21	18.05 ± 0.15 (+9.00)	26.75 ± 0.05 (+8.28)	30.40 ± 0.05 (+6.54)
Soft Teacher [51]	ICCV21	20.46 ± 0.39 (+11.41)	30.74 ± 0.08 (+12.27)	34.04 ± 0.14 (+10.18)
PseCo [19]	ECCV22	22.43 ± 0.36 (+13.38)	32.50 ± 0.08 (+14.03)	36.06 ± 0.24 (+12.2)
DINO [55] (Supervised)	-	18.00 ± 0.21	29.50 ± 0.16	35.00 ± 0.12
Omni-DETR [46] (DINO)	CVPR22	27.60 (+9.60)	37.70(+8.20)	41.30 (+6.30)
Semi-DETR [56] (DINO)	CVPR23	30.5 ± 0.30 (+12.50)	40.10 ± 0.15 (+10.6)	43.5 ± 0.10 (+8.5)
Sparse Semi-DETR	-	**30.9 ± 0.23** (+12.90)	**40.8 ± 0.12** (+11.30)	**44.3 ± 0.01** (+9.30)

Table 1. Comparing Sparse Semi-DETR with other approaches on COCO-Partial setting. The results are the average across all five folds. Under the COCO-partial setting, FCOS serves as the baseline for one-stage detectors, Faster RCNN for two-stage detectors, and DINO for transformer-based end-to-end detectors.

Methods	Labels	COCO-Partial		
		AP_S	AP_M	AP_L
Semi-DETR [56]	1%	13.6	31.2	40.8
	5%	23.0	43.1	53.7
	10%	25.2	46.8	58.0
Sparse Semi-DETR	1%	14.8	32.5	41.4
	5%	23.9	44.2	54.2
	10%	26.9	48.0	59.6

Table 2. Experimental results on COCO-partial settings for small, medium, and large objects. The results shown are the average across all five folds. We reproduce Semi-DETR results using their source code.

Methods	VOC12	
	AP_{50}	$AP_{50:95}$
FCOS [42] (Supervised)	71.36	45.52
DSL [3]	80.70	56.80
Dense Teacher [57]	79.89	55.87
Faster RCNN [35] (Supervised)	72.75	42.04
STAC [39]	77.45	44.64
HumbleTeacher [41]	80.94	53.04
Instant-Teaching [58]	79.20	50.00
DINO [55] (Supervised)	81.20	59.60
Semi-DETR [56] (DINO)	86.10	65.20
Sparse Semi-DETR	**86.30**	**65.51**

Table 3. Experimental results on Pascal VOC protocol. Here, FCOS, Faster RCNN, and DINO are the supervised baselines.

Method	COCO-Full (100%)
STAC [39] (18×)	39.5 $\xrightarrow{-0.3}$ 39.2
Unbiased Teacher (9×)	40.2 $\xrightarrow{+1.1}$ 41.3
SoftTeacher [51] (24×)	40.9 $\xrightarrow{+3.6}$ 44.5
DSL [3] (12×)	40.2 $\xrightarrow{+3.6}$ 43.8
Dense Teacher [57] (18×)	41.2 $\xrightarrow{+3.6}$ 46.1
PseCo (24×)	41.0 $\xrightarrow{+5.1}$ 46.1
Instant-Teaching [58] (24×)	37.6 $\xrightarrow{-0.27}$ 40.2
Semi-DETR [56] (8×)	48.6 $\xrightarrow{+1.8}$ 50.4
Sparse Semi-DETR (8×)	**49.2** $\xrightarrow{+2.1}$ **51.3**

Table 4. Comparing Sparse Semi-DETR with other approaches on COCO-Full. Note that 1 denotes 30k training iterations, while an N× signifies N times 30k iterations.

effective design. Applying QR selectively to the student, teacher, or both networks, we find that employing QR on weak augmented image features F_{t1} and F_{t2}, and integrating them into the student network with the original decoder queries yields the best results. Table 5c shows that processing F_{t1} features without MLP (in [56]) improves results. Table 5d presents our study on the impact of different query variants, where we observe that QR consistently outperforms other methods. Finally, Table 5e shows that using an attentional block in the QR module is more effective than simple concatenation or cosine similarity. We provide more analysis of the Query Refinement Module supplementary document. Figure 5 compares Sparse Semi-DETR and Semi-DETR. Sparse Semi-DETR, processing fewer but refined queries, demonstrates lower duplication rates in this training approach.

Effect of Reliable Pseudo-Label Filtering Module. In our analysis of the Pseudo-Label Filtering Module, we examine the impact of various parameters. Table 6a shows that a smaller m results in lower performance due to the inclusion of poor-quality labels. Performance improves with a moderately larger m due to enhanced auxiliary loss. However, excessively large m values trigger NMS, negatively impacting performance. Additionally, in Table 6b, we analyze the selection of the k value and find that setting k to 4 yields the best performance. In Table 6c, the optimal performance for σ is achieved at 0.4; values lower than this may lead to the generation of noisy pseudo-labels, whereas higher values can decrease the number of effective pseudo-labels.

Limitation. We still have duplications when we apply the one-to-many training strategy in both stages. Table 7 illustrates that a one-to-many strategy for 120k iterations with Sparse Semi-DETR achieves a 44.6 mAP using NMS. In comparison, 60k iterations in the first stage attain a comparable 44.3 mAP without NMS. Future works can explore this aspect.

Query Refinement	Pseudo-Label Filtering	mAP	AP_{50}	AP_{75}
✗	✗	43.5	58.9	46.0
✓	✗	43.8	61.1	47.3
✗	✓	43.9	60.5	46.3
✓	✓	44.3	61.7	47.6

(a) Effect of Individual Component.

Teacher	Student	mAP	AP_{50}	AP_{75}
✗	✗	43.5	58.9	46.0
✓	✗	43.8	61.2	47.2
✗	✓	44.3	61.7	47.6
✓	✓	42.8	59.7	45.8

(b) Effect of QR on Student and Teacher module.

MLP	mAP	AP_{50}	AP_{75}
✓	43.7	60.9	46.9
✗	44.3	61.7	47.6

(c) Effect of MLP.

Method	mAP	AP_{50}	AP_{75}
Single-view Queries	43.0	59.3	46.3
Cross-view Queries	43.5	58.9	46.0
Query Refinement	44.3	61.7	47.6

(d) Effect of different variants of queries.

Method	mAP	AP_{50}	AP_{75}
Simple Concat	43.4	58.8	46.1
Cosine Similarity	43.7	60.3	46.1
Attention Module	44.3	61.7	47.6

(e) Effectiveness of Attentional module in QR.

Table 5. Ablations for the proposed Sparse Semi-DETR on COCO 10% Label dataset. **(a)** We analyze the effectiveness of each module of Sparse Semi-DETR. **(b)** We experiment with different QR combinations to identify the optimal design, applying QR selectively on the student, the teacher, or both. **(c)** We analyze the effect of using MLP layers for F_{t1} and F_{s1}. Empirical observation reveals that we do not require MLP as in [56]. **(d)** We observe the effect of the attentional block in the Query Refinement module by replacing it with simple concatenation or cosine similarity. **(e)** We vary the type of object queries fed to the decoder with the decoder's original queries. The best results are highlighted .

Figure 5. The top-row figures display Semi-DETR's detection results, and the bottom-row shows Sparse Semi-DETR's outcomes. Both networks were trained for 120k iterations using one-to-many and one-to-one strategies. Sparse Semi-DETR eliminates redundant bounding boxes in the bottom-left image and detects small objects, like knives, in the top-right image, as indicated by white arrows.

m (k=4)	mAP
2	44.0
4	44.1
6	44.3
8	44.2

(a) Effect of m.

k (m=6)	mAP
2	43.8
4	44.3
6	44.1
8	43.9

(b) Effect of k.

σ	mAP
0.2	43.2
0.3	43.8
0.4	44.3
0.5	44.0

(c) Effect of σ.

Table 6. Ablations for the Reliable Pseudo-Label Filtering Module on COCO 10% Label dataset. We study the impact of various parameters as augmented ground truth (m), top pseudo-labels selection (k) and filtering threshold (σ).

Iterations	40k	60k	80k	100k	120k
mAP	43.5	44.3	44.0	43.8	44.6
without-NMS	Y	Y	Y	Y	N

Table 7. Evaluating One-to-Many Assignment Strategy.

7. Conclusion

In conclusion, we successfully address the inherent limitations of DETR-based semi-supervised object detection frameworks by introducing Sparse Semi-DETR. This novel solution effectively tackles overlapping predictions and the detection of small objects. Sparse Semi-DETR incorporates a Query Refinement Module to enhance object query quality, mainly benefiting the detection of small and partially obscured objects. Besides, it also introduces a Reliable Pseudo-Label Filtering Module to filter out low-quality pseudo-labels selectively, thereby enhancing overall detection accuracy and consistency with the remaining high-quality labels. Our method outperforms existing SSOD approaches, with extensive experiments demonstrating its effectiveness.

Ethical considerations. We study semi-supervised models, and agree that standard ethical considerations for visual recognition are applicable to our work.

Acknowledgements

This work was in parts supported by the EU Horizon Europe Framework under grant agreements 101135724 (LUMINOUS) and 101092312 (AIRISE) and by the Federal Ministry of Education and Research of Germany (BMBF) under grant 01QE2227C (HERON).

References

[1] Ting Cai and Kirthevasan Kandasamy. Active cost-aware labeling of streaming data. In *International Conference on Artificial Intelligence and Statistics*, pages 9117–9136. PMLR, 2023. 1

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, pages 213–229. Springer, 2020. 2, 3

[3] Binghui Chen, Pengyu Li, Xiang Chen, Biao Wang, Lei Zhang, and Xian-Sheng Hua. Dense learning based semi-supervised object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4805–4814, 2022. 1, 2, 3, 7

[4] Zhe Chen, Jing Zhang, and Dacheng Tao. Recurrent glimpse-based decoder for detection with transformer. *CoRR*, abs/2112.04632, 2021. 3

[5] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1601–1610, 2020. 3

[6] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015. 5

[7] Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. You only look at one sequence: Rethinking transformer in vision through object detection. *CoRR*, abs/2106.00666, 2021. 3

[8] Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fast convergence of DETR with spatially modulated co-attention. *CoRR*, abs/2101.07448, 2021. 3

[9] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 1, 2

[10] Arushi Goel, Yunlong Jiao, and Jordan Massiah. Pars: Pseudo-label aware robust sample selection for learning with noisy labels. *arXiv preprint arXiv:2201.10836*, 2022. 2

[11] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1780–1789, 2020. 4

[12] Qiushan Guo, Yao Mu, Jianyu Chen, Tianqi Wang, Yizhou Yu, and Ping Luo. Scale-equivalent distillation for semi-supervised object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14502–14511, 2022. 1, 2

[13] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[14] Shu Hu, Chun-Hao Liu, Jayanta Dutta, Ming-Ching Chang, Siwei Lyu, and Naveen Ramakrishnan. Pseudoprop: Robust pseudo-label generation for semi-supervised object detection in autonomous driving systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4390–4398, 2022. 1

[15] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *Neural Information Processing Systems*, 2019. 1, 2

[16] Ding Jia, Yuhui Yuan, Hao He, Xiao pei Wu, Haojun Yu, Weihong Lin, Lei huan Sun, Chao Zhang, and Hanhua Hu. Detrs with hybrid matching. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19702–19712, 2022. 2

[17] Purbayan Kar, Vishal Chudasama, Naoyuki Onoe, and Pankaj Wasnik. Revisiting class imbalance for end-to-end semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4569–4578, 2023. 2

[18] 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. 2, 5

[19] Gang Li, Xiang Li, Yujie Wang, Yichao Wu, Ding Liang, and Shanshan Zhang. Pseco: Pseudo labeling and consistency training for semi-supervised object detection. In *Computer Vision – ECCV 2022*, pages 457–472, Cham, 2022. Springer Nature Switzerland. 1, 2, 3, 7

[20] Jiachen Li, Haiming Gang, Hengbo Ma, Masayoshi Tomizuka, and Chiho Choi. Important object identification with semi-supervised learning for autonomous driving. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2913–2919. IEEE, 2022. 1

[21] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 5

[22] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *CoRR*, abs/1612.03144, 2016. 2

[23] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017. 2

[24] Fanfan Liu, Haoran Wei, Wenzhe Zhao, Guozhen Li, Jingquan Peng, and Zihao Li. Wb-detr: Transformer-based detector without backbone. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2959–2967, 2021. 3

[25] 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. *CoRR*, abs/2201.12329, 2022. 2

[26] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg.

Ssd: Single shot multibox detector. In *Computer Vision – ECCV 2016*, pages 21–37, Cham, 2016. Springer International Publishing. 2

[27] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 1, 3

[28] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors, 2022. 1, 2, 7

[29] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional DETR for fast training convergence. *CoRR*, abs/2108.06152, 2021. 3

[30] Yuxiang Nie, Chaowei Fang, Lechao Cheng, Liang Lin, and Guanbin Li. Adapting object size variance and class imbalance for semi-supervised object detection. In *AAAI Conference on Artificial Intelligence*, 2023. 1

[31] Ramanpreet Singh Pahwa, Richard Chang, Wang Jie, Xu Xun, Oo Zaw Min, Foo Chuan Sheng, Chong Ser Choong, and Vempati Srinivasa Rao. Automated detection and segmentation of hbms in 3d x-ray images using semi-supervised deep learning. In *2022 IEEE 72nd Electronic Components and Technology Conference (ECTC)*, pages 1890–1897, 2022. 1

[32] Luca Piras, Ludovico Boratto, and Guilherme Ramos. Evaluating the prediction bias induced by label imbalance in multi-label classification. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, page 3368–3372, New York, NY, USA, 2021. Association for Computing Machinery. 2

[33] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018. 1

[34] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, Los Alamitos, CA, USA, 2016. IEEE Computer Society. 2

[35] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017. 1, 2, 7

[36] Byungseok Roh, Jaewoong Shin, Wuhyun Shin, and Saehoon Kim. Sparse DETR: efficient end-to-end object detection with learnable sparsity. *CoRR*, abs/2111.14330, 2021. 2

[37] Claudio Filipi Gonçalves Dos Santos and João Paulo Papa. Avoiding overfitting: A survey on regularization methods for convolutional neural networks. *ACM Comput. Surv.*, 54(10s), 2022. 2

[38] Tahira Shehzadi, Khurram Azeem Hashmi, Didier Stricker, and Muhammad Zeshan Afzal. Object detection with transformers: A review, 2023. 2

[39] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *CoRR*, abs/2005.04757, 2020. 1, 3, 7

[40] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. Sparse R-CNN: end-to-end object detection with learnable proposals. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14454–14463. Computer Vision Foundation / IEEE, 2021. 3

[41] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. *CoRR*, abs/2106.10456, 2021. 3, 7

[42] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: fully convolutional one-stage object detection. *CoRR*, abs/1904.01355, 2019. 1, 2, 7

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 3

[44] Dong Wang, Yuan Zhang, Kexin Zhang, and Liwei Wang. Focalmix: Semi-supervised learning for 3d medical image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3951–3960, 2020. 1

[45] Kuo Wang, Yuxiang Nie, Chaowei Fang, Chengzhi Han, Xuewen Wu, Xiaohui Wang Wang, Liang Lin, Fan Zhou, and Guanbin Li. Double-check soft teacher for semi-supervised object detection. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1430–1436. International Joint Conferences on Artificial Intelligence Organization, 2022. Main Track. 1

[46] Pei Wang, Zhaowei Cai, Hao Yang, Gurumurthy Swaminathan, Nuno Vasconcelos, Bernt Schiele, and Stefano Soatto. Omni-detr: Omni-supervised object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9367–9376, 2022. 2, 3, 7

[47] Tao Wang, Li Yuan, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. Pnp-detr: Towards efficient visual analysis with transformers. *CoRR*, abs/2109.07036, 2021. 3

[48] Wen Wang, Yang Cao, Jing Zhang, and Dacheng Tao. FP-DETR: Detection transformer advanced by fully pre-training. In *International Conference on Learning Representations*, 2022. 3

[49] X. Wang, X. Yang, S. Zhang, Y. Li, L. Feng, S. Fang, C. Lyu, K. Chen, and W. Zhang. Consistent-teacher: Towards reducing inconsistent pseudo-targets in semi-supervised object detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3240–3249, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 1, 2

[50] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2020. 1

[51] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. *CoRR*, abs/2106.09018, 2021. 1, 2, 3, 5, 7

[52] Qize Yang, Xihan Wei, Biao Wang, Xian-Sheng Hua, and Lei Zhang. Interactive self-training with mean teachers for semi-supervised object detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5937–5946, 2021. 1, 2

[53] Xin Yu, Peng Dai, Wenbo Li, Lan Ma, Jiajun Shen, Jia Li, and Xiaojuan Qi. Towards efficient and scale-robust ultra-high-definition image demoiréing. In *European Conference on Computer Vision*, pages 646–662. Springer, 2022. 4

[54] Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ArXiv*, abs/1710.09412, 2017. 3

[55] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3, 7

[56] Jiacheng Zhang, Xiangru Lin, Wei Zhang, Kuo Wang, Xiao Tan, Junyu Han, Errui Ding, Jingdong Wang, and Guanbin Li. Semi-detr: Semi-supervised object detection with detection transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23809–23818, 2023. 2, 3, 5, 7, 8

[57] 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, 2022. 3, 7

[58] Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. *CoRR*, abs/2103.11402, 2021. 1, 2, 3, 7

[59] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable {detr}: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. 2, 3