Towards Efficient Semi-Supervised Object Detection with Detection Transformer

Jiacheng Zhang, Jiaming Li, Xiangru Lin, Wei Zhang, Xiao Tan, Hongbo Gao,
Jingdong Wang, *IEEE Fellow* and Guanbin Li, *IEEE Member*

Abstract—Semi-supervised object detection (SSOD) mitigates the annotation burden in object detection by leveraging unlabeled data, providing a scalable solution for modern perception systems. Concurrently, detection transformers (DETRs) have emerged as a popular end-to-end framework, offering advantages such as non-maximum suppression (NMS)-free inference. However, existing SSOD methods are predominantly designed for conventional detectors, leaving the exploration of DETR-based SSOD largely uncharted. This paper presents a systematic study to bridge this gap. We begin by identifying two principal obstacles in semi-supervised DETR training: (1) the inherent one-to-one assignment mechanism of DETRs is highly sensitive to noisy pseudo-labels, which impedes training efficiency; and (2) the query-based decoder architecture complicates the design of an effective consistency regularization scheme, limiting further performance gains. To address these challenges, we propose Semi-DETR++, a novel framework for efficient SSOD with DETRs. Our approach introduces a stage-wise hybrid matching strategy that enhances robustness to noisy pseudo-labels by synergistically combining one-to-many and one-to-one assignments while preserving NMS-free inference. Furthermore, based on our observation of the unique layer-wise decoding behavior in DETRs, we develop a simple yet effective re-decode query consistency training method to regularize the decoder. Extensive experiments demonstrate that Semi-DETR++ enables more efficient semi-supervised learning across various DETR architectures, outperforming existing methods by significant margins. The proposed components are also flexible and versatile, showing superior generalization by readily extending to semi-supervised segmentation tasks. Code is available at https://github.com/JCZ404/Semi-DETR.

Index Terms—Object Detection, Detection Transformer, DETR, Semi-Supervised Learning

1 Introduction

Object detection (OD) is a fundamental computer vision task that aims to predict the bounding boxes and class category of objects within an image, which has broad applications, such as autonomous driving and object tracking. In the past decade, the prosperity of deep learning [22] has led to significant advancements in this field [64], [75], [63], [47]. However, the performance of these advanced detectors heavily relies on the availability of accurately annotated datasets [48], [22], which incurs expensive annotation costs. To mitigate this reliance on labeled data, semi-supervised object detection (SSOD) [72], [53], [84] has emerged as a promising paradigm, leveraging abundant unlabeled data through semi-supervised learning (SSL) techniques [71], [5], [4] to enhance detector performance.

The current state-of-the-art methods in SSOD are predominantly built upon the conventional CNN-based detectors like Faster R-CNN [64] and FCOS [75]. These detectors typically incorporate a series of hand-crafted components, including rule-based label assignment strategies [28], [63], [65], [75], [95], [26] and non-maximum suppression (NMS) for post-processing [8]. These components complicate the overall semi-supervised detection pipeline and hinder the practical deployment. In contrast, Detection Transformers (DETRs) [11], [57], [50], [41], [92] have recently emerged as a popular alternative to the traditional object detectors. With the unique designs such as sparse queries and Hungarian matching, DETRs eliminate the need for many manual components like anchor generation and NMS, resulting in a significantly simplified and efficient end-to-end detection pipeline.

Despite the prevalence of DETRs, their potential with semi-supervised learning remains largely unexplored. This work aims to bridge this gap by systematically investigating SSOD with DETR-based detectors. Our analysis shows that a naive application of existing SSOD frameworks leads to suboptimal performance (see Section 3.2), and we trace this failure to two root causes - the very mechanisms underpinning DETRs' success, the one-to-one Hungarian matching, and the query-based decoder, also present unique challenges for standard semi-supervised techniques. First, the one-to-one assignment is highly sensitive to noise; when presented with imperfect pseudo-labels, it tends to incorrectly reject high-quality candidate detections as negatives, thereby misguiding the learning process. Second, the attention-driven, dynamic nature of the query-based decoder makes it challenging to establish stable query correspondences across different augmented inputs, which is crucial for effective consistency regular-

- *Jiacheng Zhang, Jiaming Li, Xiangru Lin, and Guanbin Li are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China. Guanbin Li is also with Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou, 510006, China, and also with Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: zhangjch58@mail2.sysu.edu.cn, lijm48@mail2.sysu.edu.cn, liguanbin@mail.sysu.edu.cn).*
- *Wei Zhang, Xiao Tan, and Jingdong Wang are with Baidu Inc., Shenzhen 518066, China (e-mail: zhangwei99@baidu.com, tanxiao01@baidu.com, welleast@gmail.com).*
- *Hongbo Gao is with the Department of Automation, School of Information Science and Technology, and Institute of Advanced Technology, University of Science and Technology of China, Hefei 230026, China, and also with Nanyang Technological University, 639798, Singapore (e-mail: ghb48@ustc.edu.cn).*
- *Corresponding Author: Guanbin Li.*

ization. Consequently, the effectiveness of both pseudo-labeling and consistency training, the two effective modern SSL techniques, is severely undermined when applied to the DETR architecture.

To address these challenges, we propose Semi-DETR++, a novel framework for efficient end-to-end semi-supervised object detection with DETRs. Our solution is built upon two key components designed to overcome the specific limitations identified. First, to mitigate the optimization inefficiency caused by noisy pseudo-labels, we introduce a Stage-wise Hybrid Matching (SHM) strategy. This approach divides the training process into two distinct phases. During the early stage, we employ a soft one-to-many assignment to salvage high-quality candidate detections that might be suppressed by label noise, while simultaneously down-weighting unreliable proposals. As training progresses and pseudo-labels become more reliable, we seamlessly transition to the standard one-to-one Hungarian matching in the later stage. This design progressively eliminates duplicate predictions, thereby preserving the end-to-end inference nature of DETRs while dramatically improving training robustness and efficiency in the presence of noise. Second, we propose a simple yet effective Re-decode Query Consistency (RQC) scheme to enable effective consistency regularization for the query-based decoder. Our approach is based on the observation that as object queries propagate through the decoder layers, they gradually aggregate increasingly discriminative features and exhibit a strong local decoding behavior, binding with specific image regions. With this insight, the RQC scheme feeds the dense predictions and the corresponding decoded features from the final decoder layer through the teacher and student decoders once more. These refined features act as an implicit guidance, prompting the queries to aggregate context from the same salient regions. The consistency loss is then applied between the outputs of this ”re-decoding” process to encourage the learning of the decoding locality. This elegantly circumvents the need for explicit query correspondence matching, providing a natural and effective regularization mechanism for the query-based architecture. By integrating these two dedicated designs, Semi-DETR++ enables highly efficient semi-supervised learning for DETR-based detectors. Extensive experiments on standard benchmarks like MS-COCO and PASCAL VOC demonstrate the superior effectiveness of our approach, establishing a new state-of-the-art performance in SSOD and surpassing previous methods by a substantial margin. Our method is termed Semi-DETR++ as it represents a significant extension and enhancement of our preliminary work presented in CVPR 2023 [94]. In summary, our principal contributions are as follows:

1) We present Semi-DETR++, a novel framework for semi-supervised object detection specifically designed for DETR-based architectures. To the best of our knowledge, this is the first systematic study to explore and identify the core challenges of applying DETRs in the SSOD task.
2) We propose a *stage-wise hybrid matching* strategy that dynamically transitions from a noise-robust one-to-many assignment to the standard one-to-one assignment. This approach effectively mitigates the training inefficiency caused by noisy pseudo-labels in the DETR framework while preserving its end-to-end inference nature.
3) We introduce a *re-decode query consistency* (RQC), a simple yet effective regularization method tailored for the query-based decoder. This innovation enables efficient consistency training by leveraging the model’s own decoding outputs, eliminating the need for complex query selection and correspondence matching.
4) Extensive experiments demonstrate that Semi-DETR++ establishes a new state-of-the-art across various SSOD settings on the MS-COCO and PASCAL VOC benchmarks, outperforming previous methods by significant margins.

Beyond our preliminary conference version [94], this journal extension offers the following new contributions:

1) A systematic, in-depth analysis of the challenges in developing DETR-based semi-supervised object detection methods, particularly in pseudo-labeling and consistency training (Sec. 3.2). This study not only completes our methodology but also provides valuable insights for the community.
2) A novel and more effective consistency scheme, the *Re-decode Query Consistency* (RCQ), which eliminates the manual and complex construction of consistent queries required in our prior work. A comprehensive analysis (Sec. 5.6) validates its superiority is further provided to validate its superiority.
3) An extension of the Semi-DETR++ framework to segmentation tasks, demonstrating its versatility by achieving strong performance on both semi-supervised instance and semantic segmentation, and underscoring the superior generalization capacity of our proposed approach.

2 Related Work

2.1 Object Detection

Modern object detection has undergone substantial progress, largely driven by advances in deep learning. Early breakthroughs were dominated by Convolutional Neural Network (CNN)-based architectures, which can be broadly categorized into two-stage and one-stage detectors.

Two-stage detectors, pioneered by the R-CNN family [29], [28], [65], operate through a coarse-to-fine process. They first generate a sparse set of region proposals and then perform classification and bounding-box regression on these regions. Faster R-CNN [65] is a landmark model in this category, introducing a dedicated Region Proposal Network (RPN) to efficiently generate proposals, thereby unifying the entire pipeline into a single network. In contrast, one-stage detectors [61], [62], [63], [75], [99] streamline the process by directly predicting bounding boxes and class probabilities from image features, eliminating the proposal generation step. Models like the YOLO series [61], [62], [63] and FCOS [75] exemplify this approach, which is often favored in real-time applications due to its superior inference speed. Despite their success, both two-stage and one-stage paradigms are inherently dense predictors. They rely on numerous hand-crafted components, such as rule-based label assignment strategies [95], [26], [27], [24], [46], [93], [45] and non-maximum suppression (NMS) for post-processing [8], [9]. These elements introduce complexity and can hinder optimal performance.

A paradigm shift was initiated with the introduction of the Detection Transformer (DETR) by Carion et al. [11]. By leveraging the transformer architecture [76], DETR formulates object detection as a set prediction problem. It uses a set of object queries and bipartite matching for label assignment, thereby eliminating the need for anchors and NMS. This end-to-end design sparked significant interest, leading to a series of improved variants that address initial limitations in convergence speed and performance. Notable works include Deformable DETR [100], which improves efficiency with multi-scale deformable attention, and DN-DETR [41], which accelerates convergence through denoising. Most recently, DINO [92], equipped with comprehensive enhancements, has achieved state-of-the-art detection performance.

2.2 Semi-Supervised Learning

Semi-supervised learning (SSL) is a powerful paradigm designed to improve model performance by leveraging both labeled and abundant unlabeled data. Early foundational work established key principles based on graph theory and transductive learning [6], [2], [10], as well as bootstrapping methods like self-training and co-training [82], [67], [7], [1].

In the deep learning era, SSL has been dominated by two principal techniques: Pseudo-Labeling and Consistency Training. The Pseudo-Labeling (PL) strategy [39], [82] involves generating artificial labels for unlabeled data and incorporating them into the training process as ground truth. To mitigate the confirmation bias inherent in this approach, advanced methods [89], [79], [16] often employ dynamic thresholding mechanisms to selectively generate higher-quality pseudo-labels. In contrast, Consistency Training (CT) is based on the manifold assumption, which posits that a model's predictions should remain consistent under small perturbations of the input. This is typically enforced by applying different data augmentations [91], [20], [21] to the same unlabeled sample and minimizing the divergence between the model's outputs [38], [74]. The Mean-Teacher framework [74] is a seminal work in this domain, which maintains a teacher model as an exponential moving average (EMA) of the student model, providing stable targets for consistency regularization.

Recently, a trend has emerged to synergistically combine Pseudo-Labeling and Consistency Training within unified frameworks like MixMatch [5] and FixMatch [71]. These methods often build upon the Mean-Teacher architecture to achieve state-of-the-art performance by leveraging the strengths of both techniques. For a more comprehensive overview, we refer readers to recent surveys [88].

2.3 Semi-Supervised Object Detection

Semi-supervised object detection (SSOD) seeks to enhance the performance of object detectors by leveraging large amounts of unlabeled data through semi-supervised learning methods. As a pioneering work, STAC [72] successfully adapted pseudo-labeling and consistency training strategies from SSL [82], [71], [5] to the object detection task. Subsequent research has largely focused on refining these two core techniques, primarily within the framework of two-stage and one-stage CNN-based detectors.

Advances in Pseudo-Labeling. A significant line of work aims to generate more accurate pseudo-labels. Early methods like ISMT [86], Humble-Teacher [73], and Instant-Teaching [98] employed techniques such as pseudo-label ensembling, co-rectification, and soft supervision [12] to exploit the pseudo-labels more reliably. To address class imbalance and confirmation bias in pseudo-labels, Unbiased Teacher [53] integrated the Mean-Teacher framework with Focal Loss, while ACRST [90] and CAPL [44] developed specific strategies to mitigate the negative effects of imbalanced pseudo-label distributions. Soft-Teacher [84] advanced this further with a decoupled labeling strategy and adaptive weighting. Other innovations include VCL [14], [15], which introduces virtual categories to exploit previously discarded low-confidence pseudo-labels, and Active-Teacher [58], which reformulates pseudo-labeling as an active sample selection problem to seek a more reliable pseudo-label selection strategy. TMR [56] otherwise encourages diverse pseudo-labels by leveraging representation disagreement.

Advances in Consistency Training. Another direction focuses on strengthening consistency regularization. Early efforts like CSD [34] employed simple flip augmentations for prediction consistency. MUM [35] developed more advanced image-tile augmentations for stronger regularization. Subsequent work shifted towards feature-level consistency; for instance, PseCo [43] and SED [30] introduced consistency training to promote scale-invariant learning.

Extension to One-Stage Detectors. While early SSOD methods were predominantly built on two-stage detectors, recent works [96], [78] have extended these techniques to one-stage architectures. DSL [13], [87] and Unbiased-Teacher V2 [54] made an initial comprehensive attempt, investigating adaptive pseudo-labeling and uncertainty consistency for anchor-free one-stage detectors. Other approaches like USD [19] and Dense-Teacher [97] focused on developing dense pseudo-labeling paradigms tailored for one-stage detectors such as FCOS [75] and RetinaNet [47]. We refer readers to [69] for a more comprehensive overview of SSOD methods.

In contrast to this extensive landscape of CNN-based methods, the application of SSOD to Detection Transformers (DETRs) remains relatively unexplored. Our conference paper, Semi-DETR [94], represents the first dedicated work to enable efficient semi-supervised learning with DETR architectures. A subsequent method, Sparse Semi-DETR [68], focuses on the query design for improved small-object detection. Note that concurrent to our work, Omni-DETR [77] proposed a DETR-based framework for omni-supervised learning. While Omni-DETR can be applied to the semi-supervised setting, it essentially applies a naive Mean-Teacher framework directly to DETR without addressing its unique challenges. As our study demonstrates, this direct application leads to suboptimal performance, underscoring the need for designs specifically tailored for SSOD with DETRs.

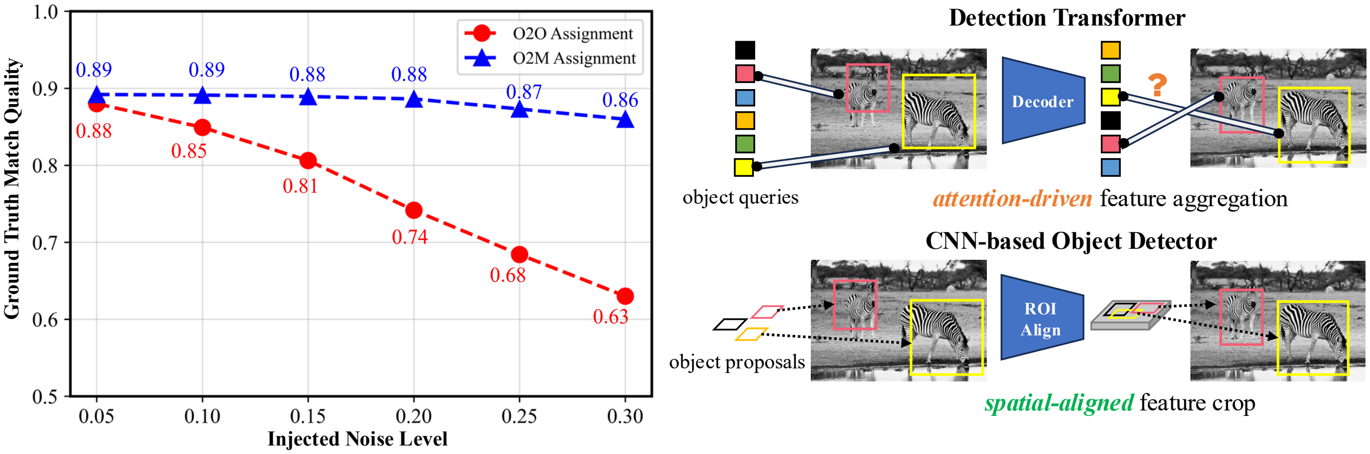


Fig. 1: **Left**: Analysis of the sensitivity of the assignment strategy. The inherent one-to-one (O2O) assignment - Hungarian matching, is more sensitive to the noise in the ground truth boxes. In contrast, the traditional one-to-many (O2M) assignment exhibits more resistance to noisy labels. **Right**: The illustration of the issue of the consistency training in DETR-based SSOD compared with the traditional CNN-based detector.

3 The Semi-Supervised Dilemma for DETRs

3.1 Preliminary

General SSOD Framework. Semi-supervised object detection (SSOD) leverages the limited labeled data alongside a large collection of unlabeled data to enhance detection performance. Formally, let $D_s = \{x_i^s, y_i^s\}_{i=1}^{N_s}$ denote the labeled dataset and $D_u = \{x_i^u\}_{i=1}^{N_u}$ the unlabeled dataset, where $N_s \ll N_u$. The annotations y^s consist of bounding box coordinates and object categories.

Modern SSOD methods predominantly resort to the Mean-Teacher framework [74]. Typically, it maintains a student model and a teacher model with identical architecture. During training, a weakly augmented version of an unlabeled image is fed to the teacher to produce pseudo-labels via filtering the dense prediction with the confidence threshold δ. A strongly augmented view of the same image is then passed to the student model and is used to train the student model with the supervision from the pseudo-labels. The overall training objective combines supervised and unsupervised losses:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_u \cdot \mathcal{L}_{\text{unsup}}, \tag{1}$$

where $\mathcal{L}_{\text{sup}}$ is computed on labeled data D_s, $\mathcal{L}_{\text{unsup}}$ is the consistency loss on unlabeled data D_u, and λ_u is a balancing weight. The student model parameters θ_s are updated via standard gradient descent, while the teacher parameters θ_t are updated as an exponential moving average (EMA) of the student:

$$\theta_s \leftarrow \theta_s - \gamma \tfrac{\partial \mathcal{L}}{\partial \theta_s}, \; \theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s, \tag{2}$$

where γ is the learning rate and α is the EMA decay rate.

End-to-End Detection Paradigm of DETR. Detection Transformer (DETR) reformulates object detection as a set prediction problem. It utilizes a set of object queries that interact with image features via cross-attention to extract and update object features. A key innovation of DETR is the use of a one-to-one (O2O) bipartite matching strategy to assign the unique best-matched prediction to each ground-truth object during training, thereby eliminating the need for handcrafted components such as anchors and non-maximum suppression (NMS). Formally, given N object query predictions, the optimal assignment $\hat{\sigma}_{\text{o2o}}$ is found by:

$$\hat{\sigma}_{\text{o2o}} = \underset{\sigma \in \xi_N}{\arg\min} \sum_{i=1}^{N} \mathcal{C}_{\text{match}}\left(\hat{y}_i, y_{\sigma(i)}\right), \tag{3}$$

where ξ_N is the set of permutations of N object query predictions, $\hat{y}_i$ and $y_{\sigma(i)}$ is the i-th ground truth and $\sigma(i)$-th object query's prediction, respectively. Note that we follow [11] to pad the ground truth with no-object to have the same length as the prediction for illustration clarity.

Generally, the matching cost $\mathcal{C}_{\text{match}}$ integrates classification, regression, and IoU costs:

$$\mathcal{C}_{\text{match}}(\hat{y}, y) = \lambda_{\text{cls}}\mathcal{C}_{\text{cls}}(\hat{y}, y) + \lambda_{\text{reg}}\mathcal{C}_{\text{reg}}(\hat{y}, y) + \lambda_{\text{iou}}\mathcal{C}_{\text{iou}}(\hat{y}, y), \tag{4}$$

where $\mathcal{C}_{\text{cls}}, \mathcal{C}_{\text{reg}}, \mathcal{C}_{\text{iou}}$ are the matching cost functions that measure the alignment between the ground truth $\hat{y}$ and the prediction y. $\lambda_{\text{cls}}, \lambda_{\text{reg}}, \lambda_{\text{iou}}$ are the correponding cost weight. This O2O assignment uniquely assigns each ground truth to its best-matching prediction, which encourages the matched prediction to become highly confident and accurate, resulting in highly distinguishable outputs that can remove duplicated predictions by simply ranking operation, thereby achieving an NMS-free inference pipeline.

3.2 Challenge of DETR-based SSOD

We delve into the two most common SSOD techniques, pseudo-labeling and consistency regularization, and reveal significant issues when applying them to DETRs.

Inefficient Training with Noisy Pseudo-Labels. We found the one-to-one assignment inherent to DETR's Hungarian matching is highly sensitive to noisy pseudo-labels, leading to significant optimization inefficiency. This sensitivity stems from the algorithm's global optimization, which selects a single, best-matching prediction for each object. When pseudo-labels are inaccurate, this process is prone to suffer from significant deviation: a low-quality prediction may be incorrectly selected as a positive sample, while a high-quality candidate is relegated to a negative sample.

This creates a fundamental optimization conflict, where the model is actively trained against its own more accurate predictions, severely hampering learning efficiency. In contrast, traditional one-to-many assignment strategies mitigate this issue by design. By allowing multiple candidates to be assigned to a single ground truth, they inherently increase robustness to label noise, as the probability of a high-quality candidate being entirely overlooked is greatly reduced.

To quantitatively validate this hypothesis, we conducted a controlled experiment comparing the sensitivity of both assignment strategies to label noise. We systematically perturb the ground-truth (GT) boxes $b_i = (x_1, y_1, x_2, y_2)$ with increasing levels of noise, following a box jittering procedure similar to [84]:

$$\begin{aligned} \hat{b}i &= (x_1 + \delta x1, y_1 + \delta_{y1}, x_2 + \delta_{x2}, y_2 + \delta_{y2}), \\ \delta &= (\delta_{x1}, \delta_{y1}, \delta_{x2}, \delta_{y2}) \sim \mathcal{N}(0,1) \odot (sw, sh, sw, sh). \end{aligned} \tag{5}$$

Here, b_i and $\hat{b}_i$ represent the original GT boxes and the perturbed version, respectively, s is the injected noise level, and w and h are the width and height of b_i. We then applied both the DETR's Hungarian matching and a traditional one-to-many assignment [24] to these noisy labels. To evaluate assignment quality, we measured the highest overlap between the assigned positive samples and the original, unperturbed ground truth boxes (termed as GT Match Quality).

It can be observed on the left of Figure 1, the quality of positive samples assigned by the one-to-one strategy degrades rapidly than the one-to-many assignment as the noise increases. This confirms that the Hungarian matcher is more vulnerable to the noisy pseudo-labels, causing it to select poor positives and, critically, leaving many high-quality predictions as negative samples. This directly creates the optimization conflict described above. Conversely, the one-to-many strategy consistently identifies positives that align well with the original ground truth, even under high noise levels, demonstrating superior robustness to localization noise and ensuring high-quality candidates receive proper supervisory signal.

Query-based Decoder Complicates the Consistency Training. Consistency regularization has proven effective in various convolution-based studies of SSOD. Its success stems from enforcing consistent predictions between different augmented inputs, such as weak and strong augmentation [34], or scale augmentation [30], [43]. This exerted consistency regularization facilitates the learning of a more discriminative representation that facilitates more effective and robust detection. However, one critical condition of such a consistency scheme is the presence of clear deterministic correspondence relationships across different augmented views, which is used to construct the pair to impose the consistency constraint. As depicted in Figure 1, the traditional CNN-based detectors like Faster-RCNN [65] operate on local regions via `ROIAlign`, which maintains the consistent spatial information throughout the processing, facilitating the establishment of corresponding region pairs for consistency regularization training [30]. However, this is not the case in detection transformers. As illustrated in the right of Figure 1, DETRs formulate the detection task as a set-prediction problem and rely on sparse object queries for adaptive attention-driven object feature aggregation. This unique design complicates the acquisition of corresponding prediction pairs necessary for consistency regularization application, as the responsible region of a particular object query constantly changes. Formally, let denote decoder queries as $\mathbf{q} = \{q_0, q_1, ..., q_{N-1}\}$ and the output of the Transformer decoder as $\mathbf{o} = \{o_0, o_1, ..., o_{N-1}\}$. F and A denote the refined image features after the transformer encoder and the attention mask derived based on the denoising task design, respectively. Then, the decode operation can be represented as:

$$\mathbf{o} = D(\mathbf{q}, F|A), \tag{6}$$

where D denotes the Transformer decoder. The reason why it is infeasible to apply the consistency regularization on DETR-based SSOD can be illustrated as,

$$\texttt{PredOrder}(\mathbf{o}) \neq \texttt{PredOrder}(\mathbf{q}), \tag{7}$$

where `PredOrder` refers to the order of each object query's prediction, which indicates the corresponding responsible regions of each query. This is also evidenced by the analysis of [41], [51]. This poses a notable challenge to implementing consistency training for the DETR-based SSOD methods. For instance, since the input query's responsible region may alter after the attention-driven feature update, it is not so straightforward to construct the consistent query pairs in different scales and then impose the consistency constraint on the model's prediction as the practice in [30].

4 Semi-DETR++

4.1 Overview

To tackle the issue when adapting the DETR for the SSOD task, we propose Semi-DETR++, which is the improved version of our conference work, Semi-DETR [94], the first DETR-based SSOD method. The overview of Semi-DETR++ is presented in Figure 2. We implement our method based on the widely used Mean-Teacher [74] framework. We develop two novel components, that is, *Stage-wise Hybrid Matching* and *Re-decode Query Consistency*, to enable more efficient and effective semi-supervised object detection with detection transformers. With these dedicated designs, Semi-DETR++ achieves end-to-end semi-supervised object detection, eliminating the annoying hand-crafted components like NMS during inference, and achieving state-of-the-art SSOD performance. Furthermore, we showcase that our method can naturally generalize to the segmentation task, exhibiting decent segmentation performance under the semi-supervised setting.

4.2 Stage-wise Hybrid Matching

To address the training inefficiency caused by inherent one-to-one bipartite matching with inaccurate pseudo-labels, we propose a stage-wise hybrid label-matching strategy, which brings the best of both one-to-many and one-to-one assignment strategies to enable efficient training while preserving the end-to-end nature of DETRs.

Given that the pseudo-labels are often noisy during early training, which may mislead the one-to-one Hungarian matching, we argue that assigning multiple potential positive candidates to pseudo-labels is more appropriate at

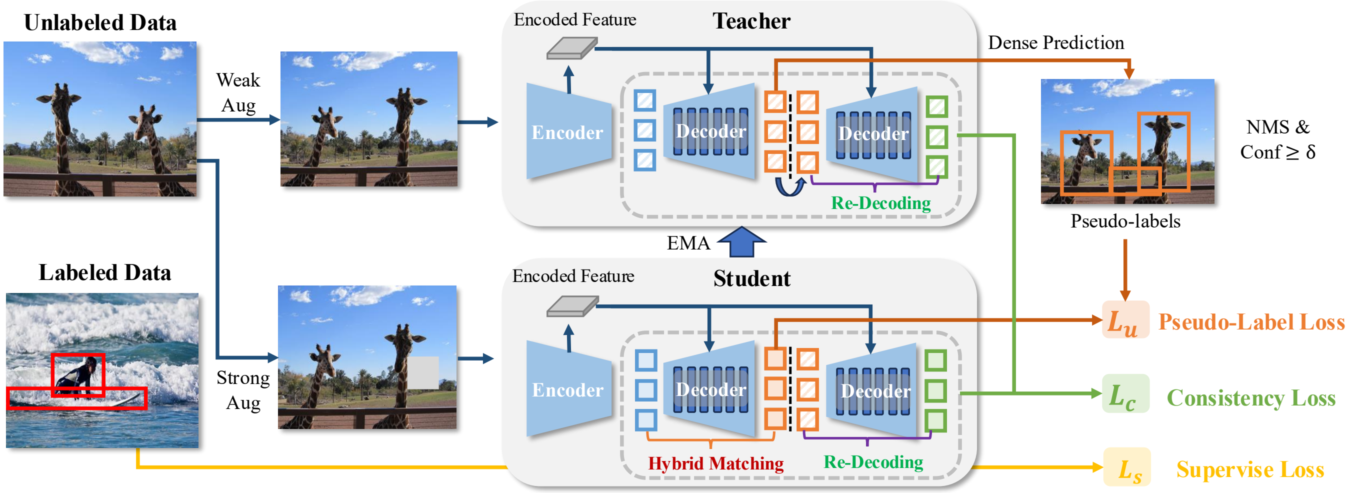


Fig. 2: Overview of Semi-DETR++. We take the teacher-student framework to implement our semi-supervised object detection with Detection Transformer (DETR). Generally, we contribute two novel designs, *Stage-wise Hybrid Matching (SHM)* and *Re-decode Query Consistency (RQC)*. The former is proposed to ease the training inefficiency in the early stage when the pseudo labels are noisy, while the latter is designed to achieve effective consistency training with DETRs. Integrating these two designs, Semi-DETR++ achieves the state-of-the-art semi-supervised object detection performance.

this period. To this end, we divide the training into two stages and employ a one-to-many assignment to mitigate the potential optimization conflicts caused by the noisy pseudo-labels in the early stage. Formally, the one-to-many assignment $\hat{\sigma}_{\text{o2m}}$ is defined as:

$$\hat{\sigma}_{\text{o2m}} = \left\{ \underset{\sigma_i \in \mathrm{C}_N^M}{\arg\min} \sum_{j=1}^{M} \mathcal{C}_{\text{match}}\left(\hat{y}_i, y_{\sigma_i(j)}\right) \right\}_{i=1}^{|\hat{\mathbf{y}}|}, \tag{8}$$

where C_N^M is the set of M combinations of N, which denotes that a subset of M proposals in N candidates is assigned to ground truth (pseudo ground truth) $\hat{y}_i$. This strategy selects M positive samples for each pseudo-label based on a matching cost. To align with DETR's design, we consider both classification and regression in the matching criteria. Specifically, we use a high-order combination of the classification score s and the IoU overlap u between predicted and ground truth bounding boxes to define the matching quality:

$$\mathcal{C}_{\text{match}}(\hat{y}_i, y_j) = -m_{ij} = -s_{ij}^{\alpha} \cdot u_{ij}^{\beta}, \tag{9}$$

where s_{ij} and u_{ij} represent the classification score and IoU overlap of the j-th prediction with respect to the i-th ground truth, respectively. The term m_{ij} indicates the overall matching quality of the j-th candidate with the i-th ground truth. The hyperparameters α and β control the effect of classification score and IoU during the assignment, with default values set to $\alpha = 1$, $\beta = 6$. We then assign multiple candidates to each ground truth based on the lowest matching cost. If a candidate is assigned to multiple ground truths, we select the ground truth with the maximum overlap. Therefore, for each positive candidate j, we can obtain its matching quality $\hat{m}_j$ with respect to the target ground truth. Since our one-to-many assignment may also include some low-quality positive ones when increasing the number of positive candidates, we then utilize the matching quality to suppress the impact of such kinds of low-quality positive samples. Concretely, we modify the classification and regression loss function to exploit these positive samples as follows:

$$\begin{aligned}\mathcal{L}_{\text{cls}}^{\text{o2m}} &= \sum_{j=1}^{N_{\text{pos}}} \left|\hat{m}_j - s_j\right|^{\gamma} \text{BCE}\left(s_j, \hat{m}_j\right) + \sum_{j=1}^{N_{\text{neg}}} s_j^{\gamma}\, \text{BCE}\left(s_j, 0\right), \\ \mathcal{L}_{\text{reg}}^{\text{o2m}} &= \sum_{j=1}^{N_{\text{pos}}} \hat{m}_j \mathcal{L}_{\text{GIoU}}\left(b_j, \hat{b}_j\right) + \sum_{j=1}^{N_{\text{pos}}} \hat{m}_j \mathcal{L}_{L1}\left(b_j, \hat{b}_j\right), \\ \mathcal{L}^{\text{o2m}} &= \mathcal{L}_{\text{cls}}^{\text{o2m}} + \mathcal{L}_{\text{reg}}^{\text{o2m}},\end{aligned} \tag{10}$$

where γ is set to 2 by default, s_j and b_j is the classification prediction and regressed bounding box location of j-th candidate. $\hat{b}_j$ and $\hat{m}$ are assigned the ground truth box and the matching quality of the target ground truth. With such an assignment strategy, the noisy pseudo-labels can be covered by multiple positive samples nearby, which reduces the risk of conflicting positive and negative proposals, and in conjunction with the modified loss function, the negative impact caused by proposals with inferior quality can also be suppressed. Although the one-to-many assignment eliminates the training inefficiency, it inevitably leads to duplicated prediction, wherein the NMS must be applied during the pseudo-label generation in the one-to-many assignment training stage. With the accuracy of pseudo-labels increasing, we transform to the original one-to-one assignment along the loss function in [11], [100] after T_1 steps, one-to-many assignment training to rescue the NMS-free property of DETRs in the late training stage. The loss is also altered into:

$$\mathcal{L}^{\text{o2o}} = \mathcal{L}_{\text{cls}}^{\text{o2o}} + \mathcal{L}_{\text{reg}}^{\text{o2o}}. \tag{11}$$

Note that in this stage, the NMS operation is retained during pseudo-label generation as it is not clear when the duplication is entirely eradicated, while the model gradually learns to remove the duplicated prediction. Ultimately, the model evolves into an NMS-free model during inference, aligning with DETR's end-to-end design.

4.3 Re-decode Consistency Training

The lack of correspondence in the responsible region between the outputs and the inputs after going through the forward decoding hinders the efficient consistency training in DETR-based SSOD.

Cross-view Consistency Training: In our previous work (Semi-DETR), we addressed the challenge of consistency training for DETRs by designing cross-view query consistency (CQC) regularization. This approach ensures local decoding behavior by leveraging the semantic context of cross-view queries. The implementation, depicted in Figure 3, follows a structured pipeline: (1) generating pseudo-consistency bounding boxes b from the teacher model's final decoder layer outputs; (2) extracting region-of-interest (ROI) features using `RoIAlign`; and (3) constructing the final cross-view queries by projecting these features and exchanging them between two distinct augmented views of the input image. These constructed cross-view queries of the student model and teacher model are then used to guide the decoding process of the teacher and student decoder, respectively, which can be formulated as follows:

$$\begin{aligned} \hat{\mathbf{q}}_{\mathbf{t}} = \texttt{MLP}(\texttt{ROIAlign}(F_t, b)),\ \hat{\mathbf{q}}_{\mathbf{s}} = \texttt{MLP}(\texttt{ROIAlign}(F_s, b)), \\ \hat{\mathbf{o}}_{\mathbf{t}} = D_t(\hat{\mathbf{q}}_{\mathbf{s}}, F_t|A), \quad \hat{\mathbf{o}}_{\mathbf{s}} = D_s(\hat{\mathbf{q}}_{\mathbf{t}}, F_s|A), \end{aligned} \tag{12}$$

where F_t and F_s are the encoded features of the teacher and student model. $\hat{\mathbf{q}}_{\mathbf{s}}$ and $\hat{\mathbf{q}}_{\mathbf{t}}$ are the constructed cross-view queries from the student and teacher models. The cross-view query consistency loss is imposed on the final decoding results of these queries. The core insight behind cross-view query consistency (CQC) training is to exploit the semantic prior to assist with query correspondence. Instead of taking the randomly initialized/constantly updated object queries to construct the consistency training pairs, it utilizes the local feature of the consistency pseudo-labels regions as the initialization of the query for consistency regularization, which constrains the decoded operation locally. The injected semantic prior enables the inherent implicit decode output to correspond, where consistency regularization is exerted to boost the semantic matching capability of the decoder.

Despite its effectiveness, CQC suffers from two primary limitations: (a) Heuristic Candidate Selection: The consistency training requires a careful selection of candidate boxes. An insufficient number leads to inadequate training, while an excess of boxes with poor semantic coverage introduces unreliable correspondences and compromises effectiveness. Although our conference version addressed this issue with a cost-based pseudo-label mining strategy, it complicates the overall pipeline. (b) Computational Overhead: The reliance on the `ROIAlign` operation introduces considerable time cost. This overhead becomes especially prominent when increasing the number of queries to enforce dense consistency regularization.

Re-decode Consistency Training: To address the issue of cross-view consistency training, we propose a novel consistency scheme that eliminates the need for tedious candidate box selection. Our approach is motivated by the insight that while object queries are global, their decoding becomes more localized as they progress through the decoder layers. To quantify this phenomenon, we define the decoding instability score (DIS) to measure the prediction order alternation of the object query across decoder layers. Formally, let the predicted objects from decoder layer i be $O = \{O_0^i, O_1^i, ..., O_{N-1}^i\}$, where N is the number of queries, and the ground-truth objects be $T = T_0, T_1, ..., T_{M-1}$, where M is the number of ground-truths. We denote the corresponding relationship between the queries' prediction and the ground truth inferred via bipartite matching with index vector $\mathrm{V^i} = \{V_0^i, V_1^i..., V_{N-1}^i\}$,

$$V_n^i = \begin{cases} m, & \text{if } O_n^i \text{ matches } T_m \\ -1, & \text{if } O_n^i \text{ matches nothing ,} \end{cases} \tag{13}$$

Then, the decoding instability score is calculated as:

$$DIS^{i-1\to i} = \sum_{j=0}^{N} \mathbb{I}\left(V_n^i \neq V_n^{i-1}\right), \tag{14}$$

We statistic the layer-wise DIS on the left of Figure 3, which clearly shows that the prediction order variation diminishes rapidly as decoding proceeds, regardless of whether it is in the early, middle, or late stages of training. In other words, *although the query has the global receptive field with attention mechanism, the decoder is learn to update the query features and make it more localized, enabling more stable and local decoding.*

Based on this observation, we introduce a re-decode query consistency (RQC) scheme for DETR-based SSOD, as illustrated in Figure 3. Specifically, we formulate the query as a combination of content and positional queries following [50], [41], [92], *i.e.* , $q_i = q_i^c + q_i^p$, where q_i^c is a learnable content query and q_i^p is the positional query initialized with the proposal box location. Given the content and positional queries of the teacher and student models, we first decode these queries as usual. Then, we use the updated candidate box from the last decoder layer of the teacher model to reinitialize the positional queries for both the teacher and student decoders. Subsequently, we conduct an additional decoding process using their corresponding updated content queries. Since each content query has already identified the best-matched positional query to aggregate features for specific objects during the first decoding process, these queries are ensured to aggregate features locally during the second decoding process. This locality enables us to construct consistent training query pairs without requiring additional matching, as follows:

$$\widetilde{\mathbf{o}}_{\mathbf{t}} = D_t(\mathbf{q}_{\mathbf{t}}^{\mathbf{c}} + \widetilde{\mathbf{q}}_{\mathbf{t}}^{\mathbf{p}}, F_t|A), \quad \widetilde{\mathbf{o}}_{\mathbf{s}} = D_s(\mathbf{q}_{\mathbf{s}}^{\mathbf{c}} + \widetilde{\mathbf{q}}_{\mathbf{t}}^{\mathbf{p}}, F_s|A) \tag{15}$$

where q_t^c and q_s^c are the learnable content query of the teacher and the student decoder after the first round decoding, $\widetilde{q}_t^p$ is the positional query derived from the teacher's dense box prediction from the last decoder layer. We then pose the consistency regularization on the classification based on the decoded queries $\widetilde{\mathbf{o}}_{\mathbf{t}}$ and $\widetilde{\mathbf{o}}_{\mathbf{s}}$:

$$\begin{aligned} w_i = \mathcal{I}(\max_{c\in[0,K]} p^t\left(y_c \mid \widetilde{o}_{t_i}\right) > \eta), \\ \mathcal{L}_{\text{unsup}}^{\text{con}} = \sum_{i=0}^{|\widetilde{o}_t|} w_i \mathcal{L}_{\text{KL}}(p^s(y_c \mid \widetilde{o}_{s_i}) \| (p^t(y_c \mid \widetilde{o}_{t_i})). \end{aligned} \tag{16}$$

where $p(\cdot|\cdot)$ is the classification head, K is the number of class, $\mathcal{L}_{\text{KL}}$ are the Kullback-Leibler (KL) divergence loss.

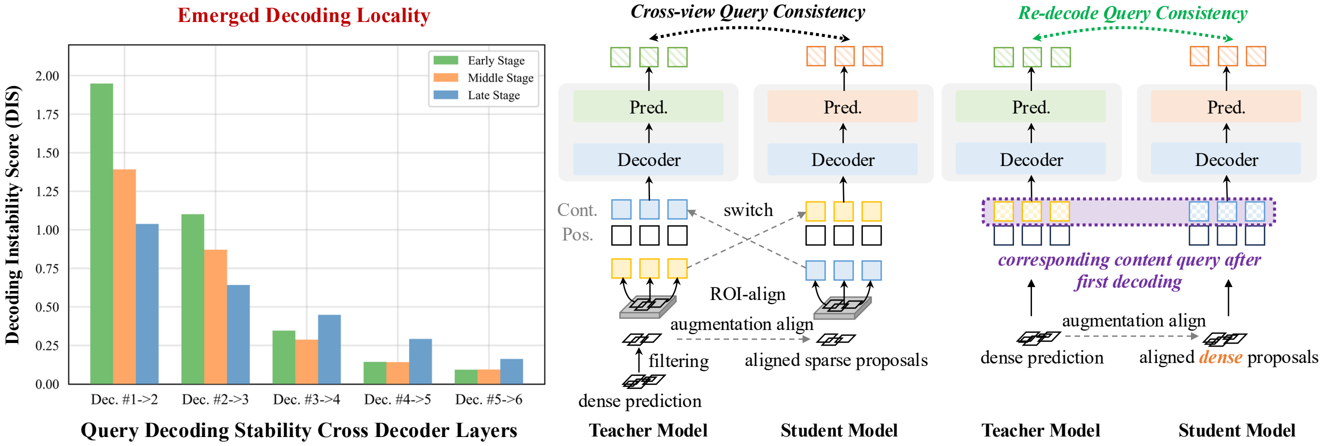


Fig. 3: **Left**: Through layer-by-layer decoding, DETRs gradually focus on the local region, resulting in more stable object queries and ground truth object correspondence. **Right**: The illustration of the cross-view query consistency and the newly proposed re-decode query consistency. 'Cont.' - content query; 'Pos.' - positional query.

We modify the KL loss by adding an extra focal term to eliminate the imbalance problem following [59]. Here, we further introduce the weight factor w_i to make the model focus more on the foreground object with the confidence threshold η. We did not add the regression loss during consistency training as we observed no obvious gains when incorporating such a term.

4.4 Training Objective

The complete training objective of our Semi-DETR++ is the combination of supervision loss and consistency loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{label}} + \mathcal{L}_{\text{unsup}}^{\text{cons}}, \tag{17}$$

where $\mathcal{L}_{\text{unsup}}^{\text{cons}}$ is the consistency loss calculated with the unlabeled data. The $\mathcal{L}_{\text{label}}$ is the label supervision loss including both labeled data and unlabeled data, and is formulated as follows:

$$\mathcal{L}_{\text{label}} = \mathbb{I}_{(t \le T_1)}(\mathcal{L}_{\text{sup}}^{\text{o2m}} + \lambda_u \mathcal{L}_{\text{unsup}}^{\text{o2m}}) + \mathbb{I}_{(t > T_1)}(\mathcal{L}_{\text{sup}}^{\text{o2o}} + \lambda_u \mathcal{L}_{\text{unsup}}^{\text{o2o}}). \tag{18}$$

Here, t is the current training step, and T_1 is the one-to-many assignment training duration. λ_u is the loss weight for the unlabeled data.

5 Experiment

5.1 Setup

Datasets and Evaluation Protocols: We evaluate our method on two standard object detection benchmarks: MS-COCO [48] and PASCAL VOC [23]. The MS-COCO dataset contains 160k labeled images across 80 categories, partitioned into `train2017` (118k), `val2017` (5k), and `test2017` (41k) splits. An additional 123k unlabeled images are provided in the `unlabeled2017` set. We follow established protocols and evaluate under two settings: (a) COCO-Partial: We treat 1%, 5%, and 10% of the `train2017` set as labeled data, using the remainder as unlabeled data. For each data split, we report the mean average precision (mAP) on the `val2017` set averaged over 5 different data folds. (b) COCO-Full: The entire `train2017` set is used as labeled data, supplemented by the `unlabeled2017` set. Performance is reported as mAP on the `val2017` set. For the PASCAL VOC benchmark, we follow previous works [13], [30] by training with VOC2007 training set as labeled data and the VOC2012 as the unlabeled data, and report AP_{50} and $\text{AP}_{50:95}$ on the VOC2007 test set.

Implementation Details: We implement Semi-DETR++ using two representative detection transformers, Deformable-DETR [100] and DINO [92], with a ResNet-50 backbone [32] pre-trained on ImageNet [22]. We use Focal Loss [47] for classification and a combination of Smooth L1 Loss and GIoU Loss [66] for regression. The number of object queries is set to 300 for Deformable-DETR and 900 for DINO. The training configurations for different settings are as follows: (a) COCO-Partial: Models are trained for 120k iterations with a batch size of 40 (5 images/GPU). The first stage of our hybrid matching lasts for 60k iterations. The labeled-to-unlabeled data ratio per iteration is 1:4, with an unsupervised loss weight $\lambda_u = 4.0$. (b) COCO-Full: Training is extended to 240k iterations with a batch size of 64 (8 images/GPU). The first hybrid matching stage lasts 180k iterations. The labeled-to-unlabeled data ratio is 1:1, with $\lambda_u = 2.0$. (c) PASCAL VOC: Models are trained for 60k iterations, with a 40k-iteration first stage. Other hyperparameters align with the COCO-Partial setting. Across all experiments, we use the Adam optimizer [36] with a learning rate of 0.0015 and no decay. The teacher model is updated via the exponential moving average (EMA) with a momentum of 0.999. For pseudo-labeling, the confidence threshold δ is set to 0.4, and the employed strong and weak augmentation is the same with Soft-Teacher [84]. The number of positive candidates M during the O2M assignment in SHM is set to 13. The foreground confidence threshold η used to calculate the consistency loss is set to 0.15.

5.2 Comparison with State-of-the-art Methods

We present a comprehensive comparison of Semi-DETR++ against current state-of-the-art (SOTA) SSOD methods on the MS-COCO and PASCAL VOC benchmarks.

MS-COCO Benchmark As summarized in Table 1, Semi-DETR++ establishes a new state-of-the-art across all data

TABLE 1: **Quantitative comparisons in COCO benchmark**. All results are the average of all 5 folds. 'Def-DETR' denotes Deformable DETR. 'Sup Only' denotes supervised-only baseline.

Method	Reference	COCO-Partial			COCO-Full
		1%	5%	10%	100%
Two-Stage Detector					
Unbiased Teacher [53]	ICLR 2020	20.75 ± 0.12	28.27 ± 0.11	31.50 ± 0.10	40.2 → 41.3 (+1.1)
Soft-Teacher [84]	ICCV 2021	20.46 ± 0.39	30.74 ± 0.08	34.04 ± 0.14	40.9 → 44.5 (+3.6)
Active-Teacher [58]	CVPR 2022	22.20	30.07	32.58	–
DTG-SSOD [42]	NeurIPS 2022	21.27 ± 0.12	31.90 ± 0.08	35.92 ± 0.26	40.9 → 45.7 (+4.8)
PseCo [43]	ECCV 2022	22.43 ± 0.36	32.50 ± 0.08	36.06 ± 0.24	41.0 → 46.1 (+5.1)
MixTeacher [49]	CVPR 2023	25.16 ± 0.26	34.06 ± 0.13	36.72 ± 0.16	40.9 → 45.7 (+4.8)
VC Learning [15]	TPAMI 2024	23.86	32.05	34.82	–
TMR-RD-v2 [56]	WACV 2024	26.91	34.37	37.74	40.2 → 46.9 (+6.7)
Single-Stage Detector					
DSL [13]	CVPR 2022	22.03 ± 0.28	30.87 ± 0.24	36.22 ± 0.18	40.2 → 43.8 (+3.6)
Dense Teacher [97]	ECCV 2022	22.38 ± 0.31	33.01 ± 0.14	37.13 ± 0.12	41.2 → 46.1 (+3.6)
Unbiased Teacher v2 [54]	CVPR 2022	22.71 ± 0.42	30.08 ± 0.04	32.61 ± 0.03	40.2 → 44.7 (+4.6)
Consistent Teacher [78]	CVPR 2023	26.30 ± 0.32	35.70 ± 0.14	40.00 ± 0.13	40.5 → 47.7 (+7.2)
DIL [87]	TIP 2025	25.00	32.20	34.45	–
End-to-End Detector					
Deformable-DETR Supervised [100]	–	11.00 ± 0.24	23.70 ± 0.13	29.20 ± 0.11	–
Deformable-DETR SSOD	–	19.40 ± 0.31	31.10 ± 0.21	34.80 ± 0.09	–
Omi-DETR (Def-DETR) [77]	CVPR 2022	18.60	30.20	34.10	–
Semi-DETR (Def-DETR)	CVPR 2023	25.20 ± 0.23	34.50 ± 0.18	38.10 ± 0.14	–
Semi-DETR++ (Def-DETR)	–	**26.00 ± 0.28**	**35.20 ± 0.20**	**39.40 ± 0.13**	–
DINO Supervised [92]	–	18.00 ± 0.21	29.50 ± 0.16	35.00 ± 0.12	–
DINO SSOD	–	28.40 ± 0.21	38.00 ± 0.13	41.60 ± 0.11	–
Omi-DETR (DINO) [77]	CVPR 2022	27.60	37.70	41.30	–
Semi-DETR (DINO) [94]	CVPR 2023	30.50 ± 0.30	40.10 ± 0.15	43.50 ± 0.10	48.6 → 50.4 (+1.8)
Sparse Semi-DETR (DINO) [68]	CVPR 2024	30.90 ± 0.23	40.80 ± 0.12	44.30 ± 0.01	49.2 → 51.3 (+2.1)
Semi-DETR++ (DINO)	–	**31.20 ± 0.21**	**41.60 ± 0.14**	**45.00 ± 0.15**	**48.6 → 53.0 (+3.8)**)

regimes on the MS-COCO benchmark. The results can be analyzed from three perspectives:

(a) Superiority over Conventional Methods. Semi-DETR++ demonstrates substantial performance gains over leading methods built on traditional two-stage and one-stage detectors. For instance, using a Deformable DETR backbone, our method surpasses PseCo by 3.57, 2.70, and 3.34 mAP under the 1%, 5%, and 10% labeled data settings, respectively. The advantage is even more pronounced with a DINO backbone, where margins over PseCo increase to 8.77, 9.10, and 8.94 mAP. Similar substantial improvements are observed against Dense Teacher. While the latest method, Consistent-Teacher, shows a slight edge over our Deformable DETR-based model—attributable to its additional, parameter-heavy feature alignment module—Semi-DETR++ with DINO decisively outperforms it by significant margins of 4.89, 5.89, and 5.00 mAP. These results underscore the superiority of our end-to-end approach, which achieves higher performance while eliminating the handcrafted components inherent in conventional detectors.

(b) Advancements in DETR-based SSOD. Semi-DETR++ yields a significant boost over both the supervised baseline and a naive Mean-Teacher semi-supervised baseline of DETRs. When applied to DINO, it exceeds the semi-supervised baseline by 2.80, 3.60, and 3.40 mAP, confirming that a direct application is suboptimal and that our tailored design is crucial. Furthermore, Semi-DETR++ consistently outperforms our conference version, Semi-DETR, by 0.70, 1.50, and 1.50 mAP, and also surpasses the recent Sparse Semi-DETR, which specializes in small-object detection. This validates the effectiveness and generalizability of our proposed components.

(c) Effectiveness with Large-Scale Unlabeled Data. Under the COCO-Full setting, which leverages the large-scale `unlabeled2017` set, Semi-DETR++ achieves its most impressive results. It elevates the strong DINO baseline from 48.6 to 53.0 mAP, setting a new state-of-the-art record. This substantial improvement, starting from an already high baseline, powerfully demonstrates the scalability and robustness of our method when abundant unlabeled data is available.

Pascal VOC Benchmark. Semi-DETR++ demonstrates consistent and superior performance on the PASCAL VOC benchmark, as detailed in Table 2. Building upon our conference version, Semi-DETR++ significantly improves over the supervised baseline, achieving gains of 9.0 AP_{50} and 11.0 $AP_{50:95}$ with a Deformable DETR backbone (and 4.9 AP_{50} and 5.6 $AP_{50:95}$ with DINO). These results culminate in Semi-DETR++ surpassing all previous state-of-the-art SSOD methods by substantial and consistent margins across both evaluation metrics, further validating its robustness.

5.3 Extending Semi-DETR++ to Segmentation Tasks

The versatility of Semi-DETR++ is further evidenced by its strong performance on semi-supervised segmentation tasks, as shown in Table 3 and Table 4. For semi-supervised instance segmentation, our method establishes a new state-of-the-art. It outperforms the leading method, Guided Dis-

TABLE 2: **Quantitative comparisons in PASCAL VOC benchmark.** All results are the average of all 5 folds. Def-DETR denotes Deformable DETR. Sup Only denotes supervised-only baseline.

Method	AP_{50}	$AP_{50:95}$
Two-Stage Detector		
Unbiased Teacher [53]	77.37	48.69
Instant-Teaching [98]	79.20	50.00
Humble Teacher [73]	80.94	53.04
Single-Stage Detector		
DSL [13]	80.70	56.80
Dense Teacher [97]	79.89	55.87
Unbiased Teacher v2 [54]	81.29	56.87
DIL [87]	77.60	-
End-to-End Detector		
Deformable-DETR (Sup only) [100]	74.50	46.20
Deformable-DETR SSOD (Baseline)	78.90	53.40
Semi-DETR (Def-DETR) [94]	83.50	57.20
Semi-DETR++ (Def-DETR)	**84.20**	**58.30**
DINO (Sup only) [92]	81.20	59.60
DINO SSOD (Baseline)	84.30	62.20
Semi-DETR (DINO)	86.10	65.20
Semi-DETR++ (DINO)	**86.40**	**66.10**

TABLE 3: **Semi-supervised instance segmentation with Semi-DETR++** on COCO dataset. Following the SOTA method of Guided Distillation [3], we report the instance segmentation mAP on the COCO dataset.

Method	Backbone	1%	2%	5%	10%
Supervised Method					
Mask-RCNN [31]	R50	3.5	9.5	17.4	21.9
CenterMask2 [40]	R50	10.1	13.5	18.0	22.1
Mask2Former [18]	R50	13.5	20.0	26.0	30.5
Ours (Sup-Baseline)	R50	14.5	-	27.2	31.9
Semi-supervised Method					
Data distillation [60]	R50	3.8	11.8	20.4	24.2
Noisy Boundaries [81]	R50	7.7	16.3	24.9	29.2
Polite Teacher [25]	R50	18.3	22.3	26.5	30.8
Guided Distillation [3]	R50	21.5	25.3	29.9	35.0
Ours (Semi-Baseline)	R50	22.2	-	31.4	36.2
Ours	R50	**23.4**	-	**33.0**	**37.5**

TABLE 4: **Semi-supervised semantic segmentation with Semi-DETR++** on Cityscapes dataset. We report the mIoU as the performance metric.

Method	Backbone	1/16	1/8	1/4	1/2
		(186)	(372)	(744)	(1488)
ECS [70]	R50	-	67.4	70.7	72.9
CAC [37]	R50	-	69.7	72.7	-
PS-MT [52]	R50	-	67.4	70.7	72.9
U^{2}PL [80]	R50	70.6	73.0	76.3	77.2
UniMatch [85]	R50	75.0	76.8	77.5	78.6
PrevMatch [70]	R50	-	77.8	78.7	79.2
Ours	R50	**77.1**	**78.4**	**79.5**	**80.2**
CPS [17]	R101	69.8	74.3	74.6	76.8
AEL [33]	R101	74.5	75.6	77.5	79.0
PS-MT [52]	R101	-	76.9	77.6	79.1
U^{2}PL [80]	R101	74.9	76.5	78.5	79.1
PCR [83]	R101	73.4	76.3	78.4	79.1
UniMatch [85]	R101	76.6	77.9	79.2	79.5
PrevMatch [70]	R101	-	78.9	80.1	80.1
Ours	R101	**78.9**	**79.5**	**81.2**	**81.8**

tillation [3], by 1.9, 3.1, and 2.5 mask mAP under the 1%, 5%, and 10% labeled data settings, respectively. This consistent advantage highlights the transferability of our semi-supervised learning framework. This trend extends to semi-supervised semantic segmentation on the Cityscapes benchmark, where Semi-DETR++ also surpasses recent semi-supervised methods across various data regimes. The strong performance across these diverse tasks can be attributed to two factors: (1) the inherent architectural flexibility of DETR-based models, which facilitates a seamless transition from detection to segmentation, and (2) our core semi-supervised designs—the stage-wise hybrid matching and redecode query consistency—which provide a general and effective mechanism for leveraging unlabeled data beyond the object detection task.

5.4 Qualitative Comparison

We present visual comparisons of Semi-DETR++ for semi-supervised object detection under varying labeled data ratios on the COCO-partial dataset, as illustrated in Figure 5. The results demonstrate that Semi-DETR++ consistently outperforms the supervised baseline, particularly in challenging scenarios such as densely overlapped objects (*e.g.* occluded giraffes, persons on motorcycles) and partially visible objects (*e.g.* trucks). These findings underscore the effectiveness of our method in leveraging unlabeled data to enhance the base detector's performance. Furthermore, we showcase the performance of Semi-DETR++ in semi-supervised instance segmentation, as depicted in Figure 6. Compared to the state-of-the-art method [3], Semi-DETR++ achieves superior results with limited annotated data, generating more complete and precise masks. For instance, our method excels in segmenting complex objects such as flying birds and cats. These results highlight the superiority of our framework in exploiting unlabeled data to improve the performance of detection transformers across both detection and segmentation tasks.

5.5 Ablation Study

We conduct a series of ablation studies to validate the effectiveness of the proposed components in Semi-DETR++ and to determine the optimal design choices and hyperparameters. Unless otherwise specified, all experiments use DINO with a ResNet-50 backbone and are trained on 10% of the COCO labeled data.

TABLE 5: **Components ablation** of Semi-DETR++.

Stage-wise	Re-decode	mAP	AP_{50}	AP_{75}
Hybrid Matching	Query Consistency			
		41.6	58.3	45.10
✔		44.0	60.8	47.5
	✔	43.7	60.1	47.1
✔	✔	**45.0**	**62.1**	**48.6**

Ablation Study of Components. As shown in Table 5, each proposed component contributes significantly to the final performance. Starting from a naive semi-supervised baseline for DETR, the introduction of our Stage-wise Hybrid Matching (SHM) yields a substantial gain of +2.4 mAP. This confirms that mitigating optimization conflicts from noisy pseudo-labels is critical for effective semi-supervised learning with DETRs. Separately, integrating the Redecode Query

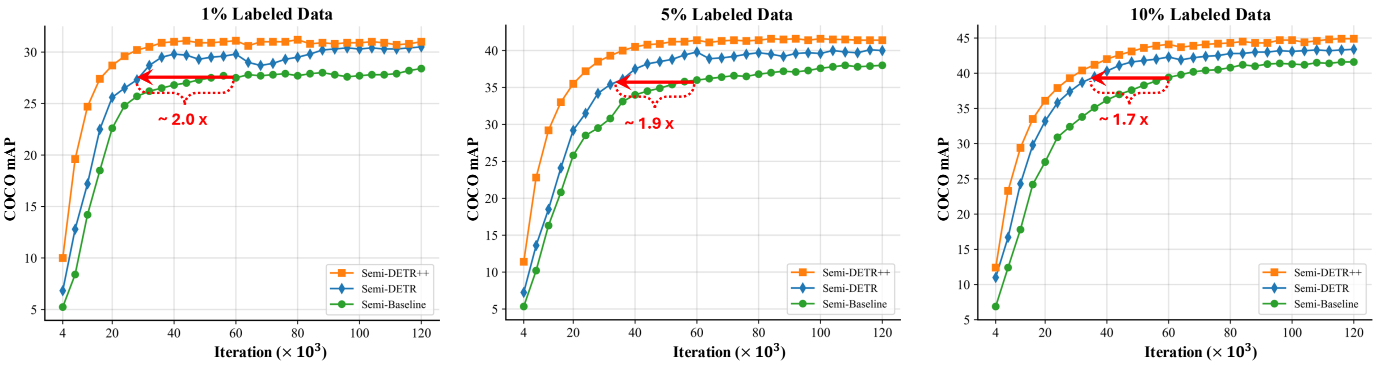


Fig. 4: **Training time evaluation curve**. (a) Semi-DETR/Semi-DETR++ employ stage-wise hybrid matching to eliminate the inefficient training caused by the noisy pseudo-labels in the early stage, improving training efficiency significantly compared to the naive semi-supervised baseline. (b) The newly introduced re-decode query consistency further boosts the convergence speed and reaches higher performance than Semi-DETR.

TABLE 6: **Hyperparameter ablation study.** (a) **PL Conf.** - confidence threshold used to filter the pseudo-labels. (b) **FG Conf.** - foreground confidence threshold used to filter the prediction to construct the consistency query. (c) **Num Pos.** - number of the positive samples assigned during O2M assignment in SHM. (d) **Unsup. Loss Weight** - loss weight of unlabeled data. (e) **O2M Iters.** - the duration of the one-to-many assignment during SHM training.

PL Conf.	mAP	FG Conf.	mAP	Num Pos.	mAP	Unsup. Loss Weight	mAP	O2M Iters. ($\times 10^4$)	mAP	NMS-Free
0.2	43.1	0.05	44.1	9	44.4	0.5	42.8	4	44.4	✓
0.3	44.3	0.10	44.7	11	44.6	1.0	43.7	6	45.0	✓
0.4	**45.0**	**0.15**	**45.0**	**13**	**45.0**	2.0	44.3	8	44.7	✓
0.5	44.5	0.20	44.5	15	44.8	**4.0**	**45.0**	10	44.5	✓
0.6	43.7	0.25	44.2	17	44.7	8.0	43.8	**12**	**45.3**	✗

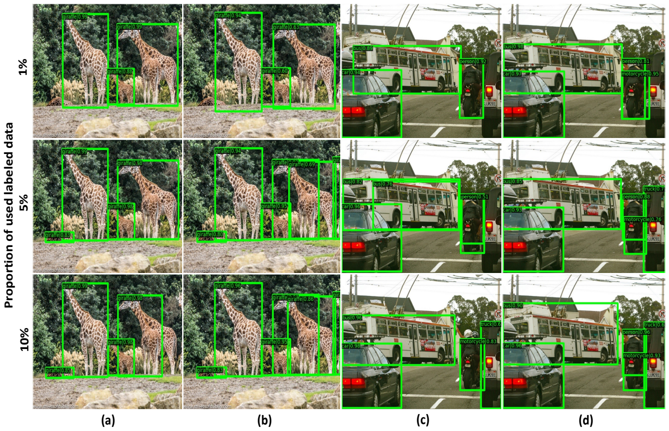


Fig. 5: A visual comparison between Semi-DETR++ (*i.e.* (b), (d)) and the baseline method (*i.e.* (a), (c)). It can be observed that Semi-DETR++ performs better under challenging cases, such as similar appearance and high occlusion, demonstrating the superiority of Semi-DETR++.

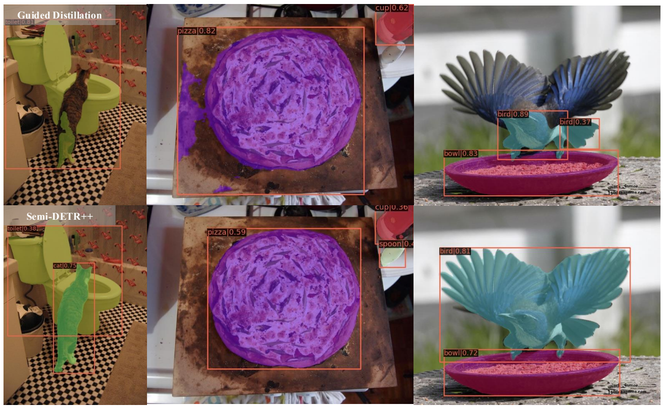


Fig. 6: A visual comparison between Semi-DETR++ and Guided Distillation on semi-supervised instance segmentation under 10% labeled data budget, where Semi-DETR++ generates more complete and precise segmentation masks.

Consistency (RQC) scheme into the baseline improves performance by +1.9 mAP. This improvement stems from the scheme's ability to enhance decoding locality and stabilize the matching process by guiding queries to focus on consistent image regions. Crucially, when combined, SHM and RQC exhibit a synergistic effect, pushing the performance to 45.0 mAP. This result not only surpasses the sum of their individual gains but also establishes a new state-of-the-art for semi-supervised object detection under the 10% COCO setting, underscoring the complementary nature of our two core designs.

More Efficient Training with Hybrid Label Matching. We evaluate the efficacy of our Stage-wise Hybrid Matching (SHM) by analyzing the training convergence curves in Figure 4. Compared to a baseline using only one-to-one (O2O) bipartite matching [11], both Semi-DETR and Semi-DETR++ — when equipped with SHM — exhibit markedly faster convergence and superior final performance. For instance, when compared with the performance of the 60k iteration of the baseline, *Semi-DETR converges over* ***1.5 times*** *faster than the naive semi-supervised manner, and reach nearly* ***2.0 times*** *under the challenging 1% labeled data setting*. This accelerated convergence stems from the one-to-many (O2M) assignment in SHM, which provides a richer supervisory signal by designating multiple positive proposals. This design effectively counteracts the instability caused by noisy pseudo-labels, leading to more robust optimization. The resulting stability, in turn, fosters the generation of more accurate pseudo-labels for subsequent training cycles. The benefit of SHM is most pronounced in low-data regimes, where the performance gap is largest. This underscores the critical role of a robust assignment strategy when pseudo-label quality is inherently volatile. Collectively, these results validate the superiority of the O2M phase in SHM for mitigating optimization conflicts and effectively harnessing noisy pseudo-labels.

More Effective Re-decode Consistency. As shown in Figure 4, our newly proposed re-decode query consistency scheme delivers substantial performance gains over our previous version. When integrated with the same stage-wise hybrid matching, Semi-DETR++ consistently outperforms our conference version (Semi-DETR), which used cross-view query consistency, across all labeled data settings (1%, 5%, and 10%) throughout the training process. The superiority of the new design stems from two key advantages. First, it eliminates the complex process of constructing a consistent query set, thereby simplifying the pipeline and improving training speed by removing the computationally expensive `RoIAlign` operation. Second, it achieves dense consistency regularization by directly leveraging all decoded query features from the initial pass. This dense supervisory signal aligns with recent trends in SSOD for traditional detectors [97], [42] and proves to be both more efficient and more effective than our previous sparse consistency approach.

Design Choice of Consistency Training. We conduct an ablation study on three key aspects of our consistency training: (1) ***How to construct the consistency queries?*** The most straightforward design is to utilize the randomly initialized query. However, this design compromises the overall performance, particularly in terms of the AP_{75} metric as shown in Table 7 (see row a). This is because the regions ultimately attended to by random queries are unpredictable due to dynamic layer-by-layer updates, leading to misaligned consistency pairs. Our conference version, Semi-DETR, addressed this with a cross-view query consistency (see row b), which uses semantic features to guide decoding towards local regions. This provides a significant performance boost by establishing a semantic prior that stabilizes query correspondence. However, it requires a special design to obtain the eligible consistency pseudo boxes, complicating the overall pipeline. In contrast, our Semi-DETR++ introduces a more elegant and effective solution: the re-decode query consistency. It uses the teacher's dense box predictions from the first decoding pass as positional queries for a second decoding step in both the teacher and student models, while retaining their original content queries (see row f). This design leverages high-quality positional guidance from the teacher to ensure local feature aggregation, eliminating the need for explicit query matching. Compared to sparsely selected queries, this dense consistency regularization is more comprehensive and effective. We also try to utilize the content queries after the first decoding pass of the student model (RD Stu.) and the teacher model (RD Tea.) to serve as the consistency query of both the teacher and the student decoder (see rows c, d). However, we found that it achieves inferior performance than the one that only the positional queries are replaced with the teacher's dense predictions while maintaining original content queries (see row f). (2)***What kind consistency loss is included?*** In our consistency training, we only include the classification loss by default. This is because we found that incorporating regression consistency loss showed no significant performance improvements, as shown in Table 7 (see rows e, f). We attribute this to the geometric augmentations (*e.g.* , shear, rotation) crucial for semantic consistency. While these transformations impose effective semantic consistency regularization by enforcing the view invariant, it may lead to the spatial displacements of the bounding boxes after the transformation, leading to counterproductive regression consistency. (3) ***Which decoder layers are selected to apply the regularization?***. We build our Semi-DETR++ upon DINO, where there are a total of 6 decoder layers. We investigate the prediction of the specific decoder layers to execute our consistency regularization and summarize the results in Table 8. The performance of our consistency training varies across decoder layers. Specifically, it shows notable robust effectiveness when applied to the later stages (layers 4-6), yielding consistently high performance with slight performance fluctuation. However, enforcing consistency on the early layers, especially the first, results in a clear performance drop. We hypothesize that this is due to the potentially unstable state of query representations in the initial decoding steps, which are not yet ready for serving as stable anchors to predict some specific regions.

TABLE 7: Ablation of the source of the consistency query, and the form of consistency loss. 'Random' - random query, 'CV' - cross-view query, 'RD' - re-decode query.

	Teacher Query	Student Query	Cls Loss	Reg Loss	mAP	AP_{50}	AP_{75}
(a)	Random	Random	✔	✔	41.4	58.2	44.4
(b)	CV	CV	✔	✔	43.8	60.8	47.5
(c)	RD Stu.	RD Stu.	✔		44.5	61.4	48.0
(d)	RD Tea.	RD Tea.	✔		44.7	61.8	48.4
(e)	RD Tea.	RD Stu.	✔	✔	44.6	61.5	48.1
(f)	RD Tea.	RD Stu.	✔		**45.0**	**62.1**	**48.6**

TABLE 8: Ablation on the decoder layer to apply the consistency. We incrementally include the early decoder layers.

	Selected Decoder Layers	mAP	AP_{50}	AP_{75}
(a)	6	**45.0**	**62.1**	48.6
(b)	5,6	44.9	61.8	**48.7**
(c)	4,5,6	44.8	61.8	48.4
(d)	3,4,5,6	44.7	61.7	48.6
(e)	2,3,4,5,6	44.7	61.6	48.4
(f)	1,2,3,4,5,6	44.5	61.4	48.0

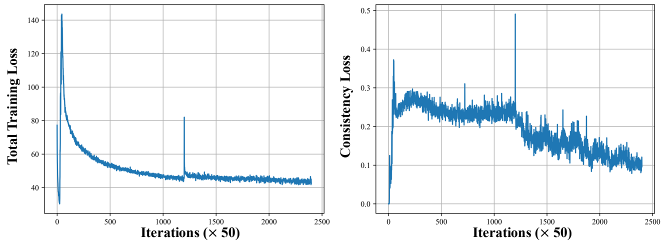


Fig. 7: **Training loss curves**. Left: The total training loss curve. Right: The consistency loss curve.

Training Convergence Stability. We visualize the training loss curve of the total loss and the consistency along in Figure 7. Overall, the total training loss is optimized smoothly, demonstrating the superior training stability of our approach despite the integration of multiple objectives. Meanwhile, the consistency objective is also optimized steadily, with the late one-to-one training stage converging faster.

Backbone Ablation. We further evaluate our approach using different backbones, including ResNet-101 [32] and Swin-Transformer Large [55], as shown in Table 9. Semi-DETR++ demonstrates consistent improvements across both architectures, confirming that its gains stem from the effective exploitation of unlabeled data and are not diminished with stronger feature extraction. Notably, with a Swin-Transformer backbone, our method sets a new state-of-the-art of 50.9 mAP using only 10% of the labeled COCO data. This result definitively shows that our framework's effectiveness is backbone-agnostic, highlighting its strong generalization capability.

TABLE 9: **Backbone ablation study.** We implement our method upon DION with ResNet-101 and Swin-Transformer Large as the backbone, respectively.

Model	1%	5%	10%
R101 Sup-only	20.0 ± 0.38	32.0± 0.12	36.7± 0.07
R101 Semi-Sup	31.3 ± 0.21	41.1 ± 0.15	43.9 ± 0.08
R101 Semi-DETR++	**32.4 ± 0.24**	**42.8 ± 0.08**	**45.7 ± 0.10**
Swin-L Sup-only	28.2± 0.29	39.7± 0.14	44.6± 0.09
Swin-L Semi-Sup	38.9 ± 0. 28	48.1 ± 0.11	49.8 ± 0.05
Swin-L Semi-DETR++	**39.8 ± 0.33**	**49.6 ± 0.12**	**50.9± 0.07**

Hyperparameter Ablation. We also conduct thorough ablation studies on some hyperparameters in our method, and the results are summarized in Table 6. Specifically, (a) ***Pseudo-label threshold*** δ: We ablate the choice of the confidence threshold used for pseudo-label selection. The pseudo-labels of unlabeled data are generated by filtering the teacher's dense prediction. We found that either too low or too high a confidence threshold causes the performance degeneration, and the threshold of 0.4 achieves the best performance. (b) ***Foreground confidence threshold*** η: In re-decode query consistency, we retain the prediction with a certain foreground probability to impose consistency regularization. We ablate such a foreground confidence threshold. It shows that the $\eta = 0.15$ yields the best performance. Setting η too low introduces many unstable queries that correspond to no meaningful region, which harms training. while an excessively high η reduces the number of reliable queries, resulting in inadequate consistency training. Note that although there is still a foreground confidence threshold that needs to be determined, it significantly simplifies the consistency query selection procedure compared with Semi-DETR, where a special module is designed to balance the precision and recall carefully. Moreover, it can be observed that it is robust within the range of 0.10 to 0.20, achieving much better performance than the previous design. (c) ***Number of positive samples*** M: We select multiple proposals for the pseudo-labels during one-to-many assignment to rescue the potentially high-quality samples. We found set $M = 13$ achieves the best result. It can be observed that further increasing the number of positive candidates leads to slight performance degradation, as it may include more negative samples than the misclassified positive samples. (d) ***Loss weight of unlabeled data*** λ_u: We analyze the impact of the unlabeled data loss weight λ_u. It shows that both excessively small and large values degrade performance. When λ_u is too small, the contribution of unlabeled data becomes negligible, causing the training to rely primarily on labeled data and limiting potential gains. Conversely, an overly large value overemphasizes the supervision from the inaccurate pseudo-labels, which can misguide the optimization process. $\lambda_u = 4$ achieves the best performance by striking a good balance between exploiting the accurately annotated data and effectively leveraging unlabeled data. (e) ***Duration of the O2M assignment*** T_1: We ablate the duration of the O2M assignment during SHM training. Note that it can achieve the best performance by employing the O2M assignment throughout the training process, which sacrifices the detection transformer's NMS-Free inference property. In comparison, it achieves decent performance with 60,000 optimization steps using O2M assignment, while retaining the NMS-free characteristic.

5.6 Mechanism Analysis

We provide additional analytical experiments to facilitate a better understanding of the mechanisms behind our approach.

How did the one-to-many assignment in SHM help the training? We propose stage-wise hybrid matching (SHM) to facilitate the efficient SSOD training by combining the merits of one-to-many assignment and one-to-one assignment. One natural question is whether other one-to-many assignment strategies work with our SHM module. To this end, we conduct an ablation study on the one-to-many (O2M) assignment strategy of SHM, investigating whether alternative O2M designs are suitable. We replace our original assignment with several common strategies: Max-IoU [65], ATSS [95], and SimOTA [27]. The performance after the O2M training stage is summarized in Table 10. A key finding is that ***not all O2M strategies perform effectively in DETR-based detectors.*** To understand why, we analyze their underlying mechanisms via several representative metrics: *Coverage* - the proportion of GT boxes assigned with at least one candidate, *Num* - the average number of candidates assigned to each ground truth, *Best Quality* - the highest overlap between the best-matched candidate and GT, *Variance* - the average spread of assigned positive candidates. We visually compare these assignment strategies in Figure 8. We observe: (a) Max-IoU assignment, designed for the dense, predefined anchor boxes of two-stage detectors [65], is ill-suited for DETRs' sparse and dynamic object queries. In conventional two-stage object detectors, the anchors with varying aspect ratios are densely and evenly distributed, which ensures sufficient positive candidates (*i.e.* anchor boxes) for each ground truth. In contrast, DETRs rely on sparse and dynamically changing object queries, where the limited set of queries learns to adapt during training, and can not guarantee sufficient IoU overlap with each ground truth box, causing proposals to cluster around a few ground truths and leaving others with inadequate coverage. (b) ATSS assignment is similarly designed for

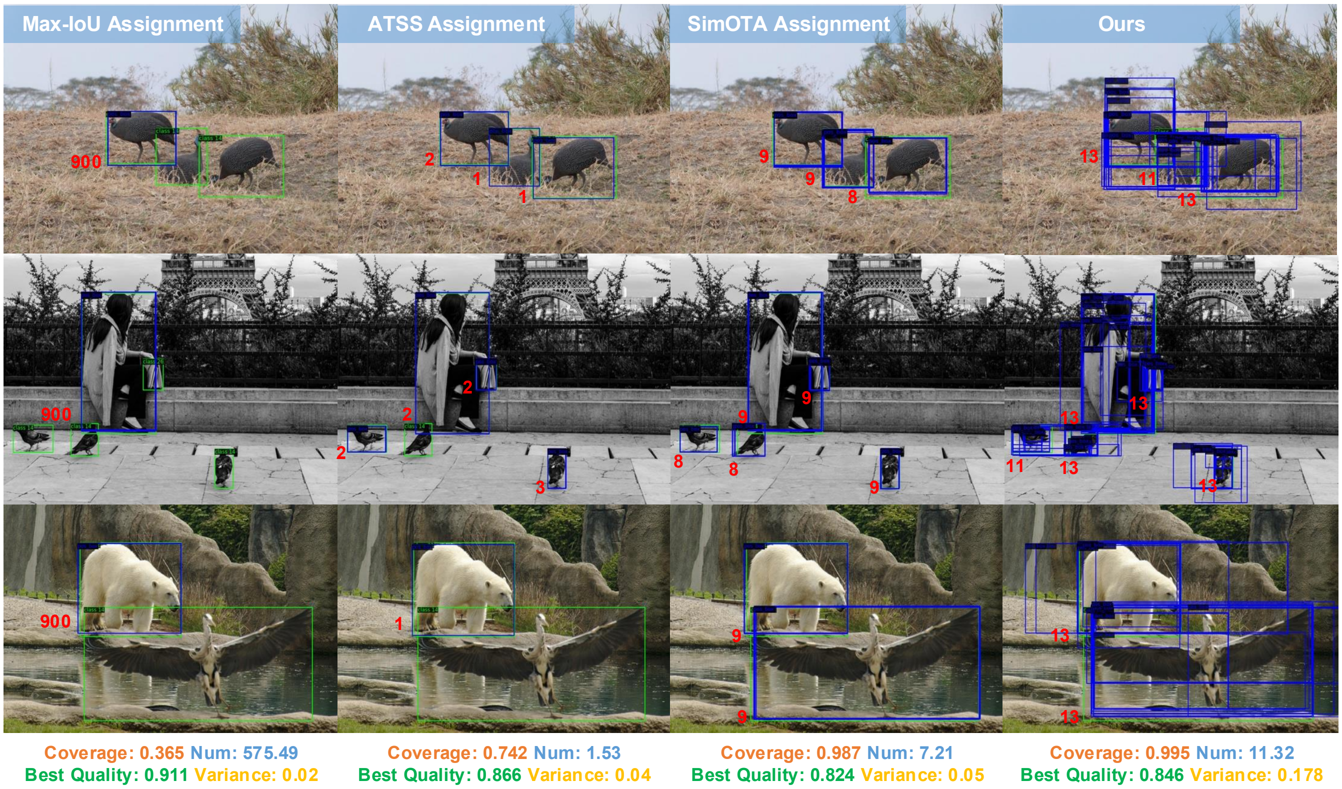

Max-IoU Assignment
ATSS Assignment
SimOTA Assignment
Ours
900
2
1
1
9
9
8
13
11
13
900
2
2
2
3
9
9
8
8
9
13
13
11
13
13
900
1
9
9
13
13
Coverage: 0.365 Num: 575.49
Best Quality: 0.911 Variance: 0.02
Coverage: 0.742 Num: 1.53
Best Quality: 0.866 Variance: 0.04
Coverage: 0.987 Num: 7.21
Best Quality: 0.824 Variance: 0.05
Coverage: 0.995 Num: 11.32
Best Quality: 0.846 Variance: 0.178

Fig. 8: **Visual analysis of one-to-many (O2M) assignment strategies** in DETRs. The red figures indicate the number of candidates assigned to each GT box. We analyze different O2M assignment characteristics via the metrics of coverage rate, positive proposal number, best-matched quality, and variance of the assigned proposal. See more details in Sec.5.6.

fixed anchors, employing the sample-wise adaptive IoU threshold $t_g = m_g + v_g$ - where m_g and v_g are the mean and standard deviation of IoU for each ground truth — to select positive candidates (*i.e.* anchor boxes) for each ground truth. However, this mechanism is incompatible with DETRs. The model's learnable object queries cause regressed boxes to cluster densely around ground truths, which artificially inflates the adaptive IoU threshold t_g. This overly stringent filtering erroneously excludes many high-quality proposals. As a result, while ATSS improves spatial coverage over other methods, the average number of positive candidates per object remains critically low (below 2), severely limiting its effectiveness. (c) SimOTA and our proposed strategy prove more robust. Both share two key principles: (1) a rank-based top-k selection to guarantee sufficient positives, and (2) a comprehensive ranking criterion that balances classification confidence and regression quality (e.g., IoU) to select diverse, high-quality candidates. However, SimOTA employs a dynamic-k matching strategy, which determines the number of positive candidates by summing the top-k largest IoUs. This approach prioritizes highly overlapping candidates (reducing assignment variance) but overlooks candidates with slight deviations. These deviated candidates are crucial when exploiting noisy pseudo-labels, as they may represent high-quality proposals when the pseudo-ground truth is inaccurate. In summary, *the SHM enhances the training efficiency by assigning sufficient positive proposals with a certain spatial deviation but retaining the decent proposal quality to each pseudo label*. These kinds of proposals deliver more robustness to the inaccurate pseudo-labels in the early stage, mitigating the potential optimization conflict and enabling more comprehensive supervision.

TABLE 10: **One-to-many (O2M) assignment analysis** in SHM.

	SHM Variants	mAP	AP_{50}	AP_{75}
(a)	O2M w/ Max-IoU [64]	11.4	15.0	12.1
(b)	O2M w/ ATSS [95]	18.7	30.5	18.9
(c)	O2M w/ SimOTA [27]	42.5	59.9	45.2
(d)	Ours	**44.1**	**61.0**	**47.5**

Why re-decode query consistency is well-suited to DETR's attention-based query mechanism? We conduct further analysis of our re-decode query consistency and reveal several key properties that make it an ideal consistency training scheme for DETR-based SSOD: (1) *Persistent Semantic Correspondence.* The quantitative experiment results about decoding instability score (DIS) shown in Figure 3 have already validated that the queries become increasingly stable and locally focused through layer-by-layer decoding, developing clear semantic bindings to particular image regions. Our method builds upon this insight and leverages this inherent property to construct consistent query pairs. To further demonstrate this persistent semantic binding, we visualize the query's attention and the corresponding prediction of the first decoding and re-decoding process in Figure 9. It can be observed that queries progressively localize correct regions during initial decoding. When these updated queries are fed through the decoder again, they maintain consistent localization with more concentrated attention. This validates our hypothesis that the queries preserve the semantic correspondence through re-decoding. The re-decode query consistency transforms the converged

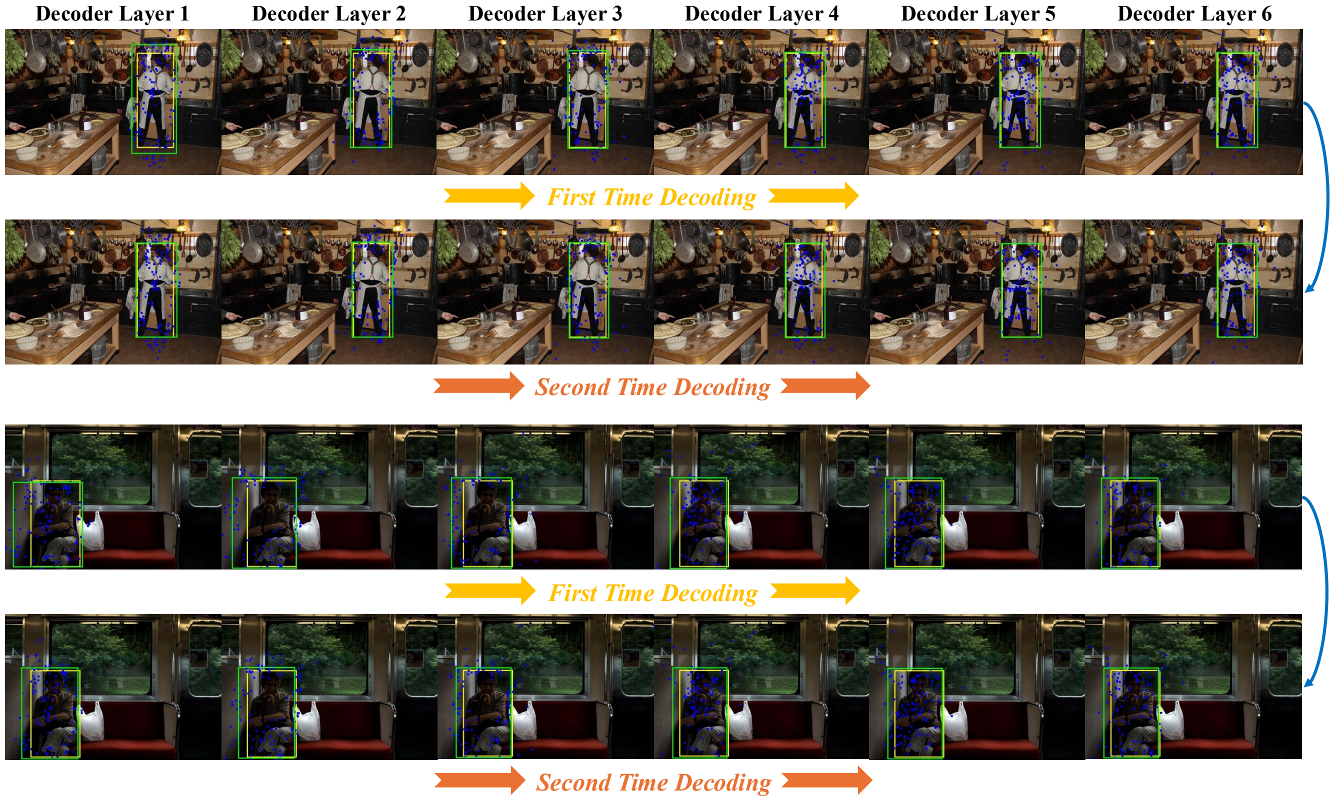


Fig. 9: **Visual illustration of the re-decode locality** in DETRs. The green rectangle is the predicted bounding box of the query, yellow rectangle is the ground truth bounding box, and the small blue squares are the deformable attention attended locations. Best viewed zoomed in.

prediction from the teacher decoder to serve as guidance for the student decoder, and then enforces the re-decoding results of the student query to be consistent with the teacher. This can be understood as a form of distillation where the student learns local decoding behavior from the teacher by aligning their re-decoding outputs. (2) *Dense Consistency Supervision.* Unlike the sparse consistency queries in our conference version, Semi-DETR++ leverages the teacher's dense predictions without selection for more comprehensive regularization. This aligns with recent findings on the advantages of dense supervision in Mean-Teacher frameworks [97], [42]. This is also demonstrated by the ablation in Table 6, where higher confidence thresholds (0.25 vs. 0.15) cause notable performance degradation. While higher thresholds ensure high foreground overlap, they discard valuable partially-overlapped predictions that contribute to effective dense supervision. (3) *Implicit Correspondence Eliminates Explicit Matching.* Our approach bypasses explicit query matching entirely by leveraging the inherent locality of decoded queries when constructing dense consistency pairs. An alternative would require Hungarian matching between query sets, which has $\mathcal{O}(n^2)$ complexity and becomes computationally expensive as query numbers increase. By exploiting the natural semantic binding of re-decode queries, our method achieves efficient consistency training without this computational bottleneck.

6 Conclusion

In this paper, we investigate the potential issue of adopting detection transformers (DETRs) for semi-supervised object detection (SSOD). We reveal that the primary challenge of DETR-based SSOD lies in the noise-vulnerable bipartite matching and the consistency training of the incompatible query-based decoding paradigm. To address these issues, we propose Semi-DETR++, the first end-to-end semi-supervised object detection approach based on detection transformers. To enhance the training efficiency, Semi-DETR++ develops a stage-wise hybrid matching strategy to combine a delicately designed one-to-many assignment strategy with the inherent bipartite matching in a stage-wise manner to resist the noisy pseudo-labels while retaining the precious end-to-end inference property of DETRs. Based on the insight that the evolved locality of query features along the layer-by-layer decoding process, a re-decode query consistency training scheme is introduced to resolve the lack of deterministic correspondence in query-based decoding and enable efficient consistency regularization with DETRs, which simply the overall pipeline compared to our conference version significantly while achieving more effective consistency regularization. Extensive experiments on COCO and PASCAL VOC benchmarks demonstrate that Semi-DETR++ outperforms existing SSOD methods by significant margins. Moreover, we show the versatility of our framework by extending it to semi-supervised segmentation tasks, including instance and semantic segmentation, highlighting its generalization capability.

REFERENCES

[1] Maria-Florina Balcan, Avrim Blum, and Ke Yang. Co-training and expansion: Towards bridging theory and practice. *NeurIPS*, 2004.

[2] Mikhail Belkin, Irina Matveeva, and Partha Niyogi. Regularization and semi-supervised learning on large graphs. In *COLT*, 2004.

[3] Tariq Berrada, Camille Couprie, Karteek Alahari, and Jakob Verbeek. Guided distillation for semi-supervised instance segmentation. In *WACV*, 2024.

[4] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019.

[5] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *NeurIPS*, 2019.

[6] Avrim Blum and Shuchi Chawla. Learning from labeled and unlabeled data using graph mincuts. 2001.

[7] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, 1998.

[8] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *ICCV*, 2017.

[9] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *ICCV*, 2017.

[10] Ulf Brefeld and Tobias Scheffer. Semi-supervised learning for structured output variables. In *ICML*, 2006.

[11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[12] Binbin Chen, Weijie Chen, Shicai Yang, Yunyi Xuan, Jie Song, Di Xie, Shiliang Pu, Mingli Song, and Yueting Zhuang. Label matching semi-supervised object detection. In *CVPR*, 2023.

[13] Binghui Chen, Pengyu Li, Xiang Chen, Biao Wang, Lei Zhang, and Xian-Sheng Hua. Dense learning based semi-supervised object detection. In *CVPR*, 2022.

[14] Changrui Chen, Kurt Debattista, and Jungong Han. Semi-supervised object detection via vc learning. In *ECCV*, 2022.

[15] Changrui Chen, Jungong Han, and Kurt Debattista. Virtual category learning: A semi-supervised learning method for dense prediction with extremely limited labels. *TPAMI*, 2024.

[16] Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. *arXiv preprint arXiv:2301.10921*, 2023.

[17] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, 2021.

[18] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022.

[19] Dayoung Chun, Seungil Lee, and Hyun Kim. Usd: Uncertainty-based one-phase learning to enhance pseudo-label reliability for semi-supervised object detection. *TMM*, 2024.

[20] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

[21] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPRW*, 2020.

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[23] 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. *IJCV*, 2015.

[24] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *ICCV*, 2021.

[25] Dominik Filipiak, Andrzej Zapała, Piotr Tempczyk, Anna Fensel, and Marek Cygan. Polite teacher: Semi-supervised instance segmentation with mutual learning and pseudo-label thresholding. *IEEE Access*, 2024.

[26] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *CVPR*, 2021.

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

[28] Ross Girshick. Fast r-cnn. In *ICCV*, 2015.

[29] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[30] Qiushan Guo, Yao Mu, Jianyu Chen, Tianqi Wang, Yizhou Yu, and Ping Luo. Scale-equivalent distillation for semi-supervised object detection. In *CVPR*, 2022.

[31] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[33] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. *NeurIPS*, 2021.

[34] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *NeurIPS*, 2019.

[35] JongMok Kim, Jooyoung Jang, Seunghyeon Seo, Jisoo Jeong, Jongkeun Na, and Nojun Kwak. Mum: Mix image tiles and unmix feature tiles for semi-supervised object detection. In *CVPR*, 2022.

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

[37] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *CVPR*, 2021.

[38] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[39] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICMLW*, 2013.

[40] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *CVPR*, 2020.

[41] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *CVPR*, 2022.

[42] Gang Li, Xiang Li, Yujie Wang, Wu Yichao, Ding Liang, and Shanshan Zhang. Dtg-ssod: Dense teacher guidance for semi-supervised object detection. 2022.

[43] 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.

[44] Hengduo Li, Zuxuan Wu, Abhinav Shrivastava, and Larry S Davis. Rethinking pseudo labels for semi-supervised object detection. In *AAAI*, 2022.

[45] Shuai Li, Chenhang He, Ruihuang Li, and Lei Zhang. A dual weighting label assignment scheme for object detection. In *CVPR*, 2022.

[46] 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. *NeurIPS*, 2020.

[47] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.

[48] 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 *ECCV*, 2014.

[49] Liang Liu, Boshen Zhang, Jiangning Zhang, Wuhao Zhang, Zhenye Gan, Guanzhong Tian, Wenbing Zhu, Yabiao Wang, and Chengjie Wang. Mixteacher: Mining promising labels with mixed scale teacher for semi-supervised object detection. In *CVPR*, 2023.

[50] 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.

[51] Shilong Liu, Tianhe Ren, Jiayu Chen, Zhaoyang Zeng, Hao Zhang, Feng Li, Hongyang Li, Jun Huang, Hang Su, Jun Zhu, et al. Detection transformer with stable matching. In *ICCV*, 2023.

[52] Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *CVPR*, 2022.

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

[54] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *CVPR*, 2022.

[55] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

[56] Seyed Mojtaba Marvasti-Zadeh, Nilanjan Ray, and Nadir Erbil-

gin. Training-based model refinement and representation disagreement for semi-supervised object detection. In *WACV*, 2024.

[57] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *ICCV*, 2021.

[58] Peng Mi, Jianghang Lin, Yiyi Zhou, Yunhang Shen, Gen Luo, Xiaoshuai Sun, Liujuan Cao, Rongrong Fu, Qiang Xu, and Rongrong Ji. Active teacher for semi-supervised object detection. In *CVPR*, 2022.

[59] Chuong H Nguyen, Thuy C Nguyen, Tuan N Tang, and Nam LH Phan. Improving object detection by label assignment distillation. In *WACV*, 2022.

[60] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *CVPR*, 2018.

[61] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[62] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, 2017.

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

[64] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015.

[65] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015.

[66] 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 *CVPR*, 2019.

[67] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. 2005.

[68] Tahira Shehzadi, Khurram Azeem Hashmi, Didier Stricker, and Muhammad Zeshan Afzal. Sparse semi-detr: Sparse learnable queries for semi-supervised object detection. In *CVPR*, 2024.

[69] Tahira Shehzadi, Didier Stricker, Muhammad Zeshan Afzal, et al. Semi-supervised object detection: A survey on progress from cnn to transformer. *arXiv preprint arXiv:2407.08460*, 2024.

[70] Wooseok Shin, Hyun Joon Park, Jin Sob Kim, and Sung Won Han. Revisiting and maximizing temporal knowledge in semi-supervised semantic segmentation. *arXiv preprint arXiv:2405.20610*, 2024.

[71] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NeurIPS*, 2020.

[72] 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.

[73] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. In *CVPR*, 2021.

[74] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 2017.

[75] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, 2019.

[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

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

[78] Xinjiang Wang, Xingyi Yang, Shilong Zhang, Yijiang Li, Litong Feng, Shijie Fang, Chengqi Lyu, Kai Chen, and Wayne Zhang. Consistent-teacher: Towards reducing inconsistent pseudo-targets in semi-supervised object detection. In *CVPR*, 2023.

[79] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Zhen Wu, and Jindong Wang. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022.

[80] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *CVPR*, 2022.

[81] Zhenyu Wang, Yali Li, and Shengjin Wang. Noisy boundaries: Lemon or lemonade for semi-supervised instance segmentation? In *CVPR*, 2022.

[82] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *CVPR*, 2020.

[83] Haiming Xu, Lingqiao Liu, Qiuchen Bian, and Zhen Yang. Semi-supervised semantic segmentation with prototype-based consistency regularization. *NeurIPS*, 2022.

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

[85] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *CVPR*, 2023.

[86] Qize Yang, Xihan Wei, Biao Wang, Xian-Sheng Hua, and Lei Zhang. Interactive self-training with mean teachers for semi-supervised object detection. In *CVPR*, 2021.

[87] Xi Yang, Penghui Li, Qiubai Zhou, Nannan Wang, and Xinbo Gao. Dense information learning based semi-supervised object detection. *IEEE TIP*, 2025.

[88] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *TKDE*, 2022.

[89] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *NeurIPS*, 2021.

[90] Fangyuan Zhang, Tianxiang Pan, and Bin Wang. Semi-supervised object detection with adaptive class-rebalancing self-training. In *AAAI*, 2022.

[91] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[92] 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.

[93] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. Varifocalnet: An iou-aware dense object detector. In *CVPR*, 2021.

[94] 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 *CVPR*, 2023.

[95] 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 *CVPR*, 2020.

[96] Yueming Zhang, Xingxu Yao, Chao Liu, Feng Chen, Xiaolin Song, Tengfei Xing, Runbo Hu, Hua Chai, Pengfei Xu, and Guoshan Zhang. S4od: Semi-supervised learning for single-stage object detection. *arXiv preprint arXiv:2204.04492*, 2022.

[97] 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.

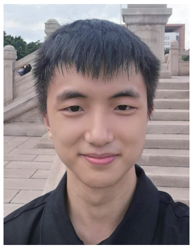
[98] Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. In *CVPR*, 2021.

[99] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[100] 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.

7 Biography Section

Jiacheng Zhang received his BS degree from Central South University, Changsha, China, in 2021. He is currently working toward an MS degree at Sun Yat-sen University. His research interests lie in deep learning and computer vision, including object detection, semi-supervised learning, and artificial intelligence-generated content.



Jiaming Li received the MS degree from Sun Yat-sen University, Guangzhou, China, in 2023. He is currently working toward a Ph.D. degree at Sun Yat-sen University. He has published several papers in top-tier academic conferences or journals, including CVPR, ICCV, and T-PAMI. His research interests lie in deep learning and computer vision, including semi-supervised learning, multi-modal learning, and object detection.

Xiangru Lin is a researcher with Sun Yat-sen University. From 2022 to 2023, he was a researcher at Baidu. He received the Ph.D. degree in computer science from The University of Hong Kong in 2021. He has published papers in venues such as CVPR, ECCV, ICCV, TPAMI, and AAAI. His research interests include computer vision and deep learning.

Wei Zhang received a Ph.D. degree from the University of Hong Kong in 2017. He is currently a senior Engineer in the Department of Computer Vision Technology at Baidu Inc. His current research interests include 2D/3D object detection, segmentation, and tracking. He is dedicated to the development of computer vision models for autonomous driving and intelligent transportation systems.

Xiao Tan Xiao Tan received the Ph.D. degree in computer vision from the University of New South Wales, Sydney, in 2014. His research interests include computer vision, pattern recognition, and image processing. He is currently with the Department of Computer Vision, Baidu as a Senior Engineer. He served as a reviewer for ICCV, CVPR, ECCV, AAAI, IJCV, and etc.

Hongbo Gao received the PH. D. degrees from Beihang University, Beijing, China, in 2016. He is currently a professor with the Department of Automation, School of Information Science and Technology, University of Science and Technology of China, Anhui Province, China, He is the author or co-author of over 100 journal papers, and he is the co-holder of 30 patent applications. He serves as an associate editor of a series of international journals including: IEEE Trans. on Neural Network and Learning System, CAAI Transactions on Intelligence Technology. His current research interests include unmanned system platform and robotics, machine learning, decision support system, intelligent driving.

Jingdong Wang (Fellow, IEEE) received the Ph.D. degree from The Hong Kong University of Science and Technology in 2007. He is currently the Chief Scientist for computer vision with Baidu. Before joining Baidu, he was a Senior Principal Researcher at Microsoft Research Asia from September 2007 to August 2021. His areas of interest include computer vision, deep learning, and multimedia search. His representative works include high-resolution network (HRNet) for generic visual recognition, transformer-based object-contextual representations (OCRNet) for semantic segmentation, discriminative regional feature integration (DRFI) for saliency detection, neighborhood graph search (NGS, SPTAG) for vector search. He has been serving/served as an Associate Editor of IEEE TPAMI, IJCV, ACM TOMM, IEEE TMM, and IEEE TCSVT, and an (senior) area chair of leading conferences in vision, multimedia, and AI, such as CVPR, ICCV, ECCV, NeurIPS, ACM MM, IJCAI, and AAAI. He will be a Program Chair for ICCV 2025. He was elected as an ACM Distinguished Member, a Fellow of IAPR, a Fellow of IEEE, and a Fellow of CAE, for his contributions to visual content understanding and retrieval.

Guanbin Li (Member, IEEE) received the Ph.D. degree from The University of Hong Kong in 2016. He is currently a full Professor with the School of Computer Science and Engineering, Sun Yat-sen University. He has authored and coauthored more than 200 papers in top-tier academic journals and conferences. His current research interests include Cross-modal visual comtent understanding and generation. He was a recipient of the ICCV 2019 Best Paper Nomination Award and the CVPR 2024 best paper candidate. He serves as an Area Chair for the conference of CVPR2024/CVPR2025/ICCV2025. He has been serving as a reviewer for numerous academic journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, International Journal of Computer Vision, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CYBERNETICS, CVPR, ICCV, ECCV, and NeurIPS.