End-to-End Semi-Supervised approach with Modulated Object Queries for Table Detection in Documents

Iqraa Ehsan[1], Tahira Shehzadi[1,2,3*], Didier Stricker[1,2,3], Muhammad Zeshan Afzal[1,2,3]

[1]Department of Computer Science, Technical University of Kaiserslautern, Germany.
[2]Mindgarage, Technical University of Kaiserslautern, Germany.
[3]German Research Institute for Artificial Intelligence (DFKI), Kaiserslautern, 67663, Germany.

*Corresponding author(s). E-mail(s): tahira.shehzadi@dfki.de;

Abstract

Table detection, a pivotal task in document analysis, aims to precisely recognize and locate tables within document images. Although deep learning has shown remarkable progress in this realm, it typically requires an extensive dataset of labeled data for proficient training. Current CNN-based semi-supervised table detection approaches use the anchor generation process and Non-Maximum Suppression (NMS) in their detection process, limiting training efficiency. Meanwhile, transformer-based semi-supervised techniques adopted a one-to-one match strategy that provides noisy pseudo-labels, limiting overall efficiency. This study presents an innovative transformer-based semi-supervised table detector. It improves the quality of pseudo-labels through a novel matching strategy combining one-to-one and one-to-many assignment techniques. This approach significantly enhances training efficiency during the early stages, ensuring superior pseudo-labels for further training. Our semi-supervised approach is comprehensively evaluated on benchmark datasets, including PubLayNet, ICADR-19, and TableBank. It achieves new state-of-the-art results, with a mAP of 95.7% and 97.9% on TableBank (word) and PubLaynet with 30% label data, marking a 7.4 and 7.6 point improvement over previous semi-supervised table detection approach, respectively. The results clearly show the superiority of our semi-supervised approach, surpassing all existing state-of-the-art methods by substantial margins. This research represents a significant advancement in semi-supervised table detection methods, offering a more efficient and accurate solution for practical document analysis tasks.

1 Introduction

A visual summary is crucial for many applications, such as document summaries and identifying graphical components in the visualization pipeline. Therefore, a crucial stage in the summary and analysis of the document will be localizing and identifying graphical elements. Tables are one of the integral elements in document analysis, adeptly consolidating crucial information into a format that maximizes spatial efficiency. Precise text extraction, especially within intricate table layouts, remains a recurring challenge. Manually extracting table objects has become impractical due to the growing number of documents. Automated methodologies [1, 2] offer dependable, impactful, and streamlined resolutions for manual tasks. OCR [3] is of paramount importance in the field of document analysis, particularly in the context of tables. It plays a crucial role in addressing the complexities associated with handling structured data. Conventional OCR[4] systems often falter when confronted with the structured format of tables, leading to misinterpretations, data loss, and formatting disparities. Early attempts [1] to address these challenges relied on external metadata embedded within documents. However, these methods proved inadequate for adapting to novel table configurations, such as borderless tables.

Recent advancements in deep learning methodologies have transformed table detection [5–8], notably through the integration of object detection concepts from computer vision. As technology advances, new challenges have emerged within the domain of table detection. A notable challenge is that training deep learning models necessitates extensive annotation of a substantial amount of data, a task that demands substantial resources. While supervised approaches [9, 10] excel on standardized benchmarks, their application in industrial contexts is according to the availability of domain-specific annotated datasets. The process of manually generating or preprocessing labeled data is not only time-consuming but also introduces a risk of inaccuracies. As a result, there's a growing demand for networks that can effectively function with limited labeled datasets.

Semi-supervised learning presents an invaluable solution precisely in situations where labeled data is limited. These techniques utilize both labeled and unlabeled data, subsequently providing pseudo-labels for the unlabeled instances. The goal of semi-supervised techniques is to improve model performance and generalization abilities by utilizing the advantages of both types of data. By doing so, it harnesses the inherent patterns and structures in the unlabeled data, enhancing the model's learning capacity and potentially improving its accuracy, all while reducing the dependency on extensive labeled datasets. Earlier, Semi-Supervised Object Detection (SSOD) [11] relies heavily on object detectors containing anchor generation process and non-maximum suppression (NMS) [12] for post-processing. Later on, DETR [13] has been adapted for semi-supervised settings, capitalizing on its inherent design which eliminates the need for NMS and efficiently utilizes both labeled and unlabeled data. The integration of semi-supervised learning in DETR presents challenges. DETR's dependence on a fixed set of object queries, established during training, may lead to inefficiencies in scenes

2

with varying object numbers, potentially causing missed detections or false positives. Furthermore, the one-to-one assignment can exacerbate errors when pseudo-labels are inaccurate. For instance, if a pseudo-label, though not entirely precise, is incorporated into the training process, the one-to-one assignment scheme may compel a predicted bounding box to closely align with this inaccurate pseudo-label.

In response to the challenges addressed, we introduce a novel semi-supervised framework based on the Teacher-Student model. Our innovative matching module integrates two training stages, utilizing one-to-many [14] and one-to-one [15] assignment strategies. The one-to-one matching ensures NMS-free end-to-end detection [12] advantages. In contrast, one-to-many matching generates higher-quality pseudo-labels for unlabeled data, contributing to a more efficiently optimized detector. In our semi-supervised learning approach, the student and teacher modules are equipped to handle both one-to-many [14] and one-to-one [15] queries. This unified processing fed object queries from the weakly augmented images to the teacher module and those from strongly augmented images to the student module. Furthermore, we implement augmented ground truth in the one-to-many matching scheme, generating more high-quality pseudo-labels.

In summary, this paper presents significant contributions in the following:

- We introduce a novel end-to-end DETR-based semi-supervised approach for table detection. The approach utilizes a dual query assignment mechanism that integrates the one-to-many matching strategy with the one-to-one matching strategy in the training phase.
- We introduce a query filtering module in a one-to-many matching strategy to provide high-quality pseudo-labels. To the best of our knowledge, this research is the first to focus on object queries in a semi-supervised setting for table detection tasks.
- We conduct a comprehensive evaluation using diverse datasets, including PubLayNet [16], ICDAR-19 [17], and TableBank [18]. Our approach surpasses both CNN-based and transformer-based semi-supervised approaches due to its superior efficiency and ability to generate high-quality pseudo-labels without using Non-maximum Suppression (NMS), resulting in a comparable performance with improved accuracy.

The subsequent sections of the paper are organized in the following manner. Section 2 provides an overview of prior research conducted in the field of detecting table objects inside document images. Section 3 explains the approach, while Section 4 presents an overview of the dataset utilized, outlines the evaluation criteria, and provides implementation details. We examine the experimental outcomes in Section 5. The ablation study is covered in Section 6. Finally, Section 7 provides a comprehensive overview of the experimental work conducted and offers insights into potential future research topics.

2 Related Work

The identification and extraction of tables in document image analysis is a crucial and significant task. Numerous studies have introduced various methods for identifying tables with diverse structures in document images. In the past, a significant number of these methodologies either depended on pre-established rules or required further meta-data input to tackle table detection tasks [20, 21]. In recent times, researchers
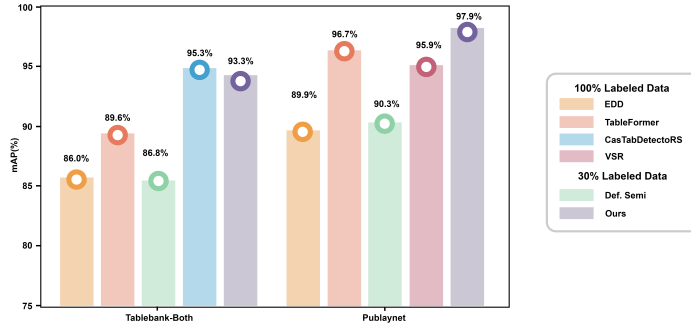
3

Fig. 1: Performance comparison of our approach with previous supervised and semi-supervised table detection approaches on TableBank-Both and Publaynet datasets. In a semi-supervised setting, we perform experiments with a 30% label data. The semi-supervised deformable transformer is referred to as the Def. Semi [19]. For an extensive summary of the results, please refer to Table 8.

have increasingly utilized statistical methodologies [22] and included deep learning techniques to augment the flexibility and adaptability of table detection systems [23–25]. This section provides an in-depth overview of these methodologies and offers insights into the strategies used for semi-supervised object detection based on Convolutional Neural Networks (CNNs).

2.1 Rule-based Strategies

Itonori et al. [26] pioneered table detection in document images by representing tables as text blocks with predefined rules. Subsequent studies by the same authors introduced a method considering both horizontal and vertical lines [27]. Pyreddy et al. [28] used tailored heuristics for tabular section detection, while Pivk et al. [29] transformed HTML formatted tables into logical forms. Hu et al. [30] proposed a method based on identifying white regions and vertically connected elements. Rule-based methods, effective for specific formats, are discussed in [31–33]. While effective, rule-based strategies may lack universal applicability, highlighting the need for more adaptable systems in the challenge of detecting tables in document images.

2.2 Learning-Based Methodologies

Cesarini et al.'s [34] supervised learning approach is intended to reliably detect and categorize table elements within document images. It entails converting an image of a document into an MXY tree structure, then recognizing and categorizing individual components as tables according to the presence of vertical and horizontal lines. To detect tables, researchers utilize Hidden Markov Models [35, 36] and a Support Vector Machine (SVM) classifier with traditional heuristics [37]. Recent advancements in deep learning methodologies have broadened their applications, extending from healthcare [38, 39], traffic analysis [40], to document analysis [41–45]. While machine learning outperforms rule-based approaches, it depends on additional information like ruling lines. In contrast, deep learning-based methodologies show higher precision and effectiveness. These

techniques can be categorized into object identification, semantic segmentation, and bottom-up approaches.

Approaches based on semantic segmentation. Table detection is framed as a segmentation problem in techniques like [46–49], leveraging existing semantic segmentation networks to create pixel-level segmentation masks. These masks are then consolidated for the ultimate table detection results, showcasing superior performance over conventional methods on benchmark datasets [16–18, 50]. In Yang et al.'s work [46], a fully convolutional network (FCN) [51] integrates linguistic and visual features, enhancing segmentation for tables and other page elements. In order to generate segmentation masks for text and table sections as well as their outlines, He et al. [47] suggested a multi-scale FCN that was further improved for the final table elements.

Inductive Approach The approaches employed in this study consider identifying tables as a problem of graph labeling, whereby the nodes of the graph represent various page components and the edges indicate the relationships between them. In their study, Li et al. [52] employed a conventional layout analysis approach to extract line regions from a given document. These line areas were further categorized into four distinct classes, including text, figure, formula, and table. This classification process was accomplished through the utilization of two CNN-CRF networks. The researchers made further predictions on the cluster that would correlate to pairs of line areas. In contrast, Holecek et al. [53] addressed the analysis of text sections by seeing them as individual nodes within a graph structure, enabling the identification and understanding of the overall document architecture. Graph-neural networks were later utilized for node-edge categorization. It is important to acknowledge that these methodologies are contingent upon certain assumptions, such as the incorporation of text line boxes as supplementary input.

Approaches Based on Object Detection. The identification and localization of tables inside document images have often been addressed as a task of object detection, initially employing R-CNN methods like those by Yi et al. [54] and Hao et al. [55], relying on heuristic rules. Advancements introduced efficient single-stage [56], [57] and two-stage [58], [59], [60], [61] detectors for document elements [62–64]. Some studies [65–67] applied image transformations for improvement, while Siddiqui et al. [68] enhanced Faster R-CNN with deformable-convolution and RoI-Pooling for geometrical modifications. The researchers Agarwal et al. [69] employed a composite network using deformable convolution to augment the performance of Cascade R-CNN. In contrast, the presented semi-supervised technique, treating detection as a set prediction task [70], eliminates stages like proposal generation and post-processing (e.g., NMS [12]), simplifying and streamlining the detection process for increased efficiency.

2.3 Object Detection with Semi-supervised

In semi-supervised object detection, two main methods emerge: consistency-based [71, 72] and pseudo-label generation-based [73–77]. Prior research [73, 74] has used diverse techniques to generate pseudo-labels, like combining prediction results or using models like SelectiveNet [78] to transfer bounding boxes from unannotated to annotated data. Research shows the superiority of self-training over pre-training, as demonstrated by STAC [79] using hard pseudo-labels. However, relying solely on initial

pseudo-label predictions may hinder model accuracy improvement. Approaches like Unbiased Teacher [80] use a two-stage framework to enhance quality via techniques like exponential moving average (EMA) and focal loss. Methods such as Mean Teacher [81], Soft Teacher [11], and Instant-Teaching [82] employ CNN-based teacher-student models for semi-supervised object detection. These methods vary in their approaches, from weighting classification losses based on teacher network outputs to employing instant pseudo-labeling schemes with extended data augmentations. However, these techniques integrate NMS into their detection process, potentially limiting training efficiency. Omni-DETR [83] and Semi-Supervised Table Detection [19] focus on stable and precise pseudo-label generation, improving the generalization capability of detectors. However, the use of one-to-one matching may introduce limitations that affect its ability to accurately recognize various objects, potentially impacting its overall performance. Our method enhances semi-supervised learning by eliminating traditional post-processing steps like NMS while improving pseudo-label quality, leading to improved model performance with limited data.

3 Method

3.1 Preliminary

The DETR model [13], which is commonly used for applications such as object detection, consists of transformer encoder-decoder network that provides object class and bounding box position[13, 84] as output. The process initiates by passing an input image through the backbone and transformer encoder [84], resulting in the extraction of image features. The decoder subsequently takes the object queries Q which is written as $\{q_1, q_2, ..., q_N\}$ and image features. The DETR model utilizes task-specific prediction heads on the embeddings of object queries ($\tilde{Q}$) at every transformer decoder layer. These prediction heads are then followed by the predictors, resulting in an output implied as $Y = \{y_1, y_2, ..., y_N\}$.

The decoder employs a self-attention mechanism on object queries, facilitating interactions between queries to gather information relevant to duplicate detection. Then, it employs cross-attention between queries and image features to extract relevant data from the image features for object detection. Lastly, it incorporates a feed-forward network that independently analyzes the queries for final detection.

DETR [13] generates a fixed number of predictions for each image, regardless of the actual number of objects present. The DETR employs a one-to-one strategy, establishing a bipartite matching between the predictions and the ground-truth objects:

$$\left(y_{\sigma(1)}, \bar{y}_1\right), \left(y_{\sigma(2)}, \bar{y}_2\right), ..., \left(y_{\sigma(N)}, \bar{y}_N\right) \tag{1}$$

In this case, the optimal permutation of N indices is denoted by $\sigma\left(\cdot\right)$, whereas the ground truth is denoted by $\bar{Y} = \{\bar{y}_1, \bar{y}_2, ..., \bar{y}_N\}$. Hence, the expression of the loss function is formulated as:

$$\mathcal{L} = \sum_{n=1}^{N} \left(y_{\sigma(n)}, \bar{y}_n\right) \tag{2}$$

where $\mathcal{L}$ represents the loss between the predicted value y and the ground truth object $\bar{y}$. The one-to-one assignment strategy in DETR matches each prediction to a unique

ground truth, reducing duplicates and simplifying object detection. This strategy simplifies the detection process and improves accuracy by eliminating the need for post-processing steps like Non-Maximum Suppression (NMS), commonly used in object detection models to handle overlapping detections. However, the one-to-one matching in DETR limits learning to a small number of positive samples per image, potentially affecting its ability to recognize various objects and impacting overall performance.

3.2 Dual Assignment Strategies

The one-to-one query assignment strategy for table detection faces challenges such as mismatches between queries and tables, leading to inefficiencies and inaccuracies in detection. It can result in unassigned queries or undetected tables, particularly in documents with a variable number of tables. This issue becomes more prominent in semi-supervised settings, where label data is limited. It needs to generate pseudo-labels for unlabeled data. However, matching these noisy pseudo-labels with object queries employing a one-to-one matching strategy results in low performance due to inaccurate pseudo-labels [19].

To address the limitations of one-to-one matching in semi-supervised setting, our approach combines both one-to-one and one-to-many matching strategies as shown in Fig. 2. This approach combines the strengths of both strategies to improve the overall efficiency of the object detection model. Within the matching strategies, two query sets are employed: $Q = \{q_1, q_2, , q_N\}$ and $\hat{Q} = \{\hat{q_1}, \hat{q_2}, , \hat{q_T}\}$. Subsequently, predictions based on Q or $\hat{Q}$ undergo either one-to-one or one-to-many matching strategies.

In one-to-one matching, the initial query set, Q, is fed to the transformer decoder layers [84], generating predictions for each layer's output. For this, bipartite matching is conducted between the pairs of predictions and ground truth across every layer. The loss is then determined using the formula:

$$\mathcal{L}_{o2o} = \sum_{l=1}^{L} \mathcal{L}_H \left(y_p, y_g\right) \tag{3}$$

Here, y_p signifies the outcomes generated by the transformer decoder layer at position l, while y_g denotes the ground-truth. The optimal assignment is efficiently calculated through the application of the Hungarian algorithm. One-to-one matching strategies are applied during both the model's training and testing phases.

In the one-to-many matching, the second set of queries, $\hat{Q}$ go through the same L transformer decoder layers, resulting in L sets of predictions. To enable one-to-many matching [14], the ground-truth was replicated K times, leading to augment ground truths $\hat{y_g} = \{y_g^1, y_g^2, , y_g^K\}$. Following this, bipartite matching is conducted between the predictions and the augmented ground truth pairs across each layer. The loss is then determined using the formula:

$$\mathcal{L}_{o2m} = \sum_{l=1}^{L} \mathcal{L}_H \left(\hat{y}_p, \hat{y}_g\right) \tag{4}$$

Here, $\hat{y}_p$ signifies the outcomes generated by the transformer decoder layer at position l. The Hungarian algorithm is employed for the proficient computation of the best assignment. The one-to-many matching strategy is exclusively applied during the model
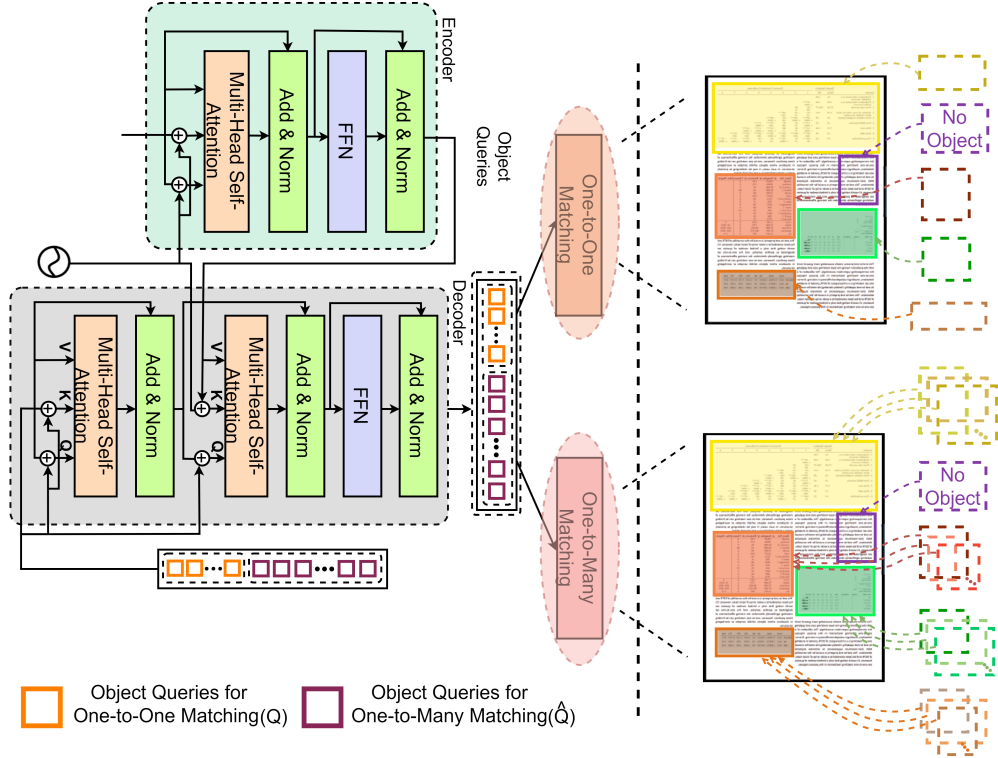
Fig. 2: Our approach contains two modules, the student and teacher module. The decoder of the student-teacher modules utilize both one-to-one and one-to-many assignment strategies. The decoder of the teacher module employs the one-to-many assignment strategy to generate high-quality pseudo-labels and improve performance with limited data. The one-to-many assignment strategy in the student module provides high-quality predictions, while the one-to-one matching strategy removes duplications.

training phase. By integrating these matching strategies, the network improves overall detection performance. The one-to-many matching strategy provides high-quality predictions, while the one-to-one matching strategy removes duplicate predictions.

3.3 Semi-Supervised Approach

In this subsection, we introduce an improved semi-supervised approach for table detection in document images. For a detailed visual representation, please refer to Fig. 3. Our approach contains two main modules: the student module and the teacher module. Both these modules contain transformer encoder-decoder network. During training, we have labeled image set $D_l = \{x_i^l, y_i^l\}_{i=1}^{N_l}$ and an unlabeled image set $D_u = \{x_i^u\}_{i=1}^{N_u}$, where N_l and N_u represents number of label and unlabel images, respectively. The student module takes both labeled and unlabeled images as input, after applying strong augmentation on the unlabeled data and a combination of strong and weak augmentation on the labeled data. The teacher's module only fed
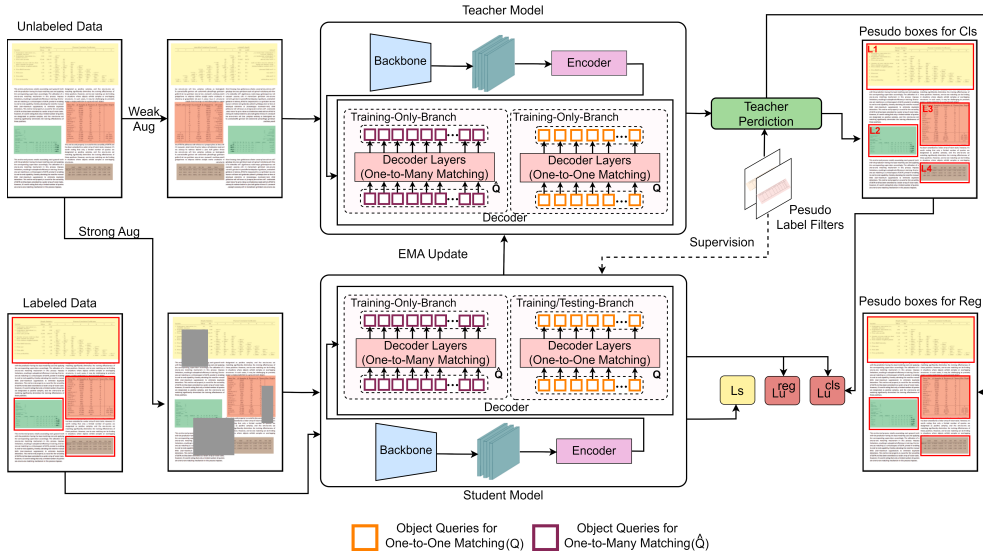
Fig. 3: **Our semi-supervised approach**: It focuses on employing one-to-many and one-to-one matching strategies in a semi-supervised setting and incorporates both labeled data and unlabeled data during training. The framework consists of two modules: student module and teacher module. The teacher module takes unlabeled images after employing weak augmentation prior and generates their pseudo-labels. The student module is designed to handle both labeled and unlabeled images, applying strong augmentation specifically to the unlabeled images. During the training process, the student module utilizes an EMA technique to consistently update the teacher module.

unlabeled data with weak augmentation as input. The teacher network is responsible for generating pseudo-labels for the unlabeled data, which the student network then uses these pseudo-labels for training. Weak augmentation for the teacher module helps in generating better pseudo-labels for unlabeled data, while strong augmentation for the student module provides better learning process. At the start of the training, student module is trained on label data fed as input and update the teacher module using Exponential Moving Average (EMA) [85]. During the training process, teacher module generates pseudo-labels for unlabel data and fed to the student module.

The decoder within these student-teacher modules utilizes two distinct matching strategies: one-to-many and one-to-one. This dual-strategy and collaborative framework between the student and teacher networks enhances the model's ability to learn from both labeled and unlabeled data, improving overall detection performance. By focusing on the quality and iterative refinement of pseudo-labels and leveraging the strengths of both matching strategies, our approach provides improved performance.

Hybrid vs traditional matching: Our primary contribution focuses on combining one-to-one and one-to-many matching strategies in a semi-supervised setting, advancing the core challenges of DETR-based Semi-Supervised table detection [19]. Previous table detection approach with limited data employs a one-to-one matching strategy. When noisy pseudo labels generated in [19] employ a one-to-one matching strategy, it impacts the overall network performance. Our approach provides multiple positive proposals

9

for each generated pseudo-label, filtering out noised pseudo-labels and selecting high-quality pseudo-labels. When fed to the student module, these high-quality pseudo-labels generated by the teacher module improve performance.

Pseudo-labels for unlabeled data: In the pseudo-labeling filtering framework, we filter out the pseudo-labels for unlabeled data generated by the teacher module. Object detectors typically produce a confidence score vector c_k for each bounding box b_k. By applying a threshold to these scores, pseudo-labels are filtered out.

In the early stages of semi-supervised table detection training, it is typical to encounter incorrect and unreliable pseudo-labels generated by the teacher module. The utilization of the one-to-one assignment approach has the potential for generating proposals that are both limited and of low quality. To improve the effectiveness of semi-supervised learning, we employ the one-to-many assignment strategy as follows:

$$\rho_{o2m} = \left\{ \underset{\sigma_i \in \chi}{argmin} \sum_{i=1}^{N} \mathcal{L}_H \left(y_i^T, y_{\sigma_i(k)}^S \right) \right\}_{i=1}^{|y^T|} \tag{5}$$

Where χ is used to represent the assignment of a subset of proposals to each pseudo label y_i^T. The term $y_{\sigma_i(k)}^S$ represents student generated prediction. In our approach, we use a one-to-many matching strategy in the teacher module to generate high-quality pseudo-labels y_i^T. The student module takes these pseudo-labels and trains the network for final predictions, employing one-to-many and one-to-one matching strategies. Achieving a one-to-one assignment involves implementing the Hungarian algorithm as follows:

$$\rho_{o2o} = \underset{\sigma \in \zeta_N}{argmin} \sum_{i=1}^{N} \mathcal{L}_H \left(y_i^T, y_{\sigma(i)}^S \right) \tag{6}$$

Here, ζ_N represents the collection of permutations involving N elements. Additionally, $\mathcal{L}_H \left(y_i^T, y_{\sigma(i)}^S \right)$ stands for the cost incurred during the matching process between the pseudo labels y^T and the prediction made by the student identified by the index $\sigma(i)$.

Losses: The teacher model generates pseudo-label for unlabeled data, while the student model is trained on both labeled images with ground-truth annotations and unlabeled images with the pseudo-labels serving as the ground-truth. This setup allows for a combined loss function, which comprises a supervised loss ($\mathcal{L}_s$) from labeled images and an unsupervised loss ($\mathcal{L}_u$) from unlabeled images, weighted by a factor ω.

$$\mathcal{L} = \mathcal{L}_s + \omega \mathcal{L}_u \tag{7}$$

Initially, the model undergoes training with a one-to-many assignment strategy, where the loss is defined by the classification loss ($\mathcal{L}_{o2m}^{cls}$) and regression loss ($\mathcal{L}_{o2m}^{reg}$).

$$\mathcal{L}_{o2m} = \mathcal{L}_{o2m}^{cls} + \mathcal{L}_{o2m}^{reg} \tag{8}$$

This approach leads to multiple positive proposals for each pseudo label, facilitating optimization and enhancing convergence speed. However, it also introduces duplicate predictions. To address this issue, the training switches to a one-to-one assignment strategy in the second stage. This transition retains the high-quality pseudo labels

obtained from the initial training stage while gradually reducing duplicate predictions. Consequently, the loss functions in this stage remain consistent with the first stage, focusing on classification and regression losses.

$$\mathcal{L}_{o2o} = \mathcal{L}_{o2o}^{cls} + \mathcal{L}_{o2o}^{reg} \tag{9}$$

A significant contribution is our matching strategy, effectively addressing core challenges in DETR-based semi-supervised table detection [19]. This leads to high-quality pseudo-labels, improving object detection accuracy. Furthermore, transitioning from one-to-one [19] to one-to-many assignment enhances the performance of semi-supervised learning. The one-to-many assignment strategy for the whole training iterations increases training efficiency. This refined approach assures a more robust and adaptable semi-supervised table detection system that improves overall performance without need NMS.

4 Experimental Setup

4.1 Dataset and Evaluation Criteria

We use TableBank [18], PubLayNet [16], and ICDAR-19 [17] for evaluating our semi-supervised approach. We utilize mean Average Precision (mAP), Average Precision at IoU threshold of 0.5 (AP^{50}), Average Precision at IoU threshold of 0.75 (AP^{75}), and Average Recall (AR_L) to assess the effectiveness of our semi-supervised approach. These evaluation metrics collectively offer a comprehensive understanding of the model's performance in object detection.

4.2 Implementation Details

We use the ResNet-50 [86] backbone, pre-trained on the ImageNet [87] dataset. We employ a one-to-one and one-to-many matching strategy for the decoder of the transformer network. We evaluate our approach on PubLayNet, ICDAR-19, and all three partitions of the TableBank dataset, utilizing 10%, 30%, and 50% of labeled data and rest as unlabeled data. The pseudo-labeling threshold is set at 0.7. Training spans 150 epochs, with a learning rate reduction by a factor of 0.1 at the 140th epoch. Strong augmentations including horizontal flips, resizing, patch removal, cropping, grayscale, and Gaussian blur are applied, alongside weak augmentation through horizontal flipping. The optimal configuration for the number of queries to the decoder input for one-to-one matching is set at 30. Additionally, 400 queries for one-to-many matching are employed, and the ground truth K is replicated six times, all contributing to the optimal outcome. Unless otherwise specified, model evaluations employ mAP (AP50:95) metrics. The weights α_1 and α_2 for balancing the classification loss (L_{cls}) and regression loss (L_{box}) are set at 2 and 5, respectively. To improve training, input image dimensions are set to 600 pixels, with the standard size of 800 pixels retained for comparative purposes with other methodologies. Fig. 4 shows the performance of our semi-supervised method applied to the PubLayNet dataset and all splits of TableBank data, illustrating the efficiency of our approach across diverse document layouts.
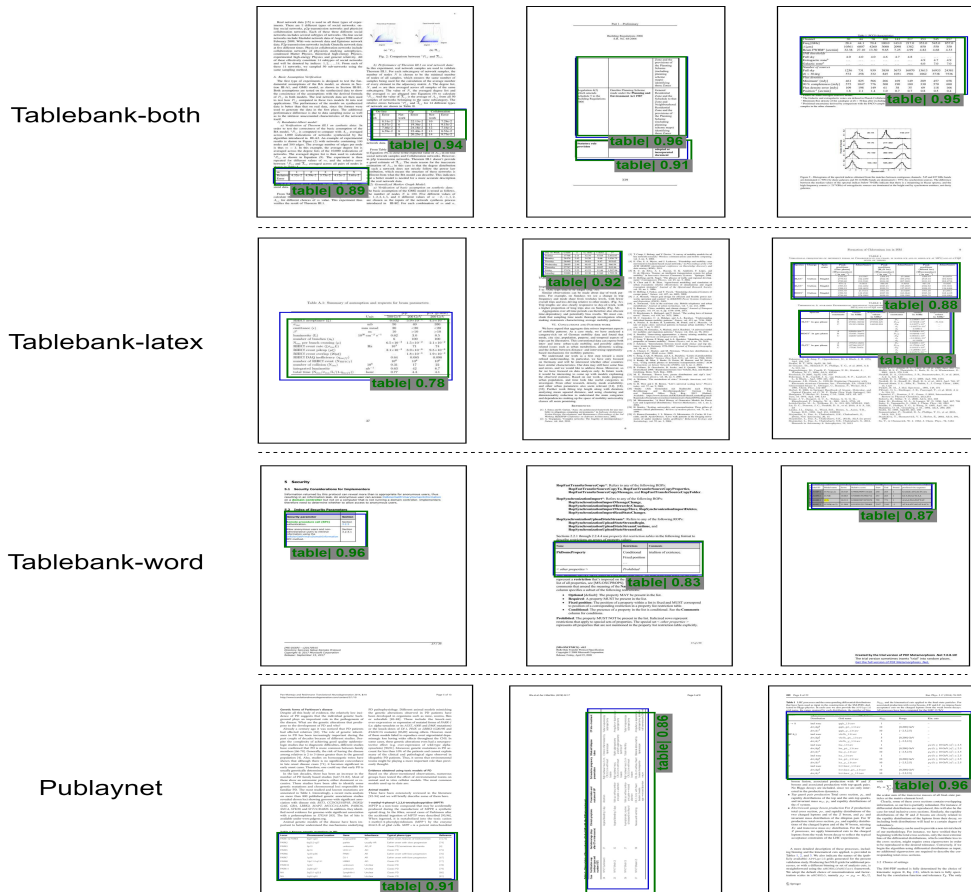
Fig. 4: Visualization of predictions of our approach on different datastets. It highlights that the inclusion of both one-to-one and one-to-many matching strategis enhance the model's accuracy. The ground truth is depicted by the blue boxes, while the green boxes display the results obtained through our approach.

5 Results and Discussion

5.1 TableBank

In this subsection, we present evaluate our semi-supervised table detection approach on all splits of the TableBank dataset [18], taking different proportions of labeled data. We also compare our transformer-based semi-supervised method, which utilizes one-to-many and one-to-one assignment strategies, against previous supervised and semi-supervised approaches. Table 1 shows the results obtained from our approach, which employs the deformable transformer [88] across the TableBank-word, TableBank-latex, and TableBank-both data splits. These experiments uses 10%, 30%, and 50% labeled data, with the remaining as unlabeled data. Notably, with 50% labeled data, the TableBank-word, TableBank-latex, and TableBank-both split achieve mAP scores of
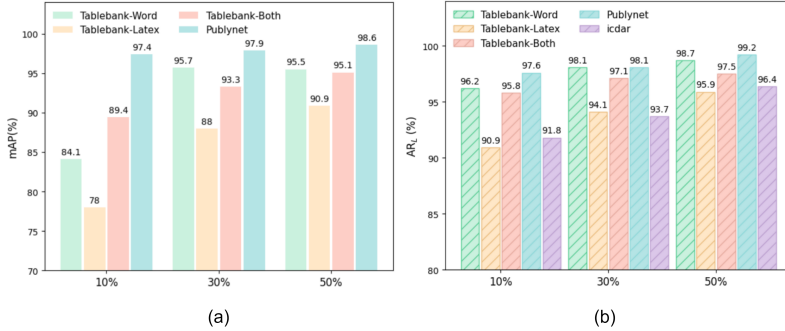
12

Fig. 5: Comparative performance analysis of our approach using the Tablebank, Publaynet, and ICDAR datasets. Experiments were conducted employing a ResNet-50 backbone with three distinct data splits: 10%, 30%, and 50%. The results are presented as follows: (a) Mean Average Precision (mAP) within the IoU threshold range of 50% to 95%, and (b) Average Recall for large objects (AR_L) within the IoU threshold range of 50% to 95%.

95.5%, 90.9%, and 95.1%, respectively, showcasing the efficiency of our semi-supervised method. Moreover, our approach shows AP^{50} of 98.5, 97.2, and 98.1 at 50% label data on TableBank-word, TableBank-latex, and TableBank-both splits, respectively. Figure 5 (a) illustrates the mAP achieved by our semi-supervised method on the PubLayNet and TableBank datasets at various percentages of labeled data. Figure 5 (b) depicts the AR_L for the PubLayNet, TableBank, and ICDAR-19 datasets, also across differing levels of labeled data. Figure 6 demonstrates the mAP at different Intersection over Union (IoU) thresholds for the TableBank-Word, TableBank-Latex, TableBank-Both, and PubLayNet datasets. Each dataset is analyzed at three percentage of labeled data: 10%, 30%, and 50%.

Dataset	Label-splits	mAP	AP^{50}	AP^{75}	AR_L
TableBank-word	10%	84.1	87.1	85.5	96.2
	30%	95.7	96.9	96.2	98.1
	50%	95.5	98.5	97.2	98.7
TableBank-latex	10%	78.0	87.5	85.5	90.9
	30%	88.0	94.3	93.1	94.1
	50%	90.9	97.2	95.9	95.9
TableBank-both	10%	89.4	94.9	93.6	95.8
	30%	93.3	96.8	95.8	97.1
	50%	95.1	98.1	97.1	97.5

Table 1: Evaluation of our semi-supervised approach across TableBank-word, TableBank-latex, and TableBank-both data divisions with varying proportions of labeled data. Here, mAP denotes the mean Average Precision within the IoU range of 0.50 to 0.95, AP^{50} signifies the Average Precision at an IoU threshold of 0.5, and AP^{75} represents the Average Precision at an IoU threshold of 0.75. Additionally, AR_L refers to the Average Recall for larger objects.

In Table 2, we present a comparison between our semi-supervised approach and earlier supervised as well as semi-supervised methods, showing results with mAP scores
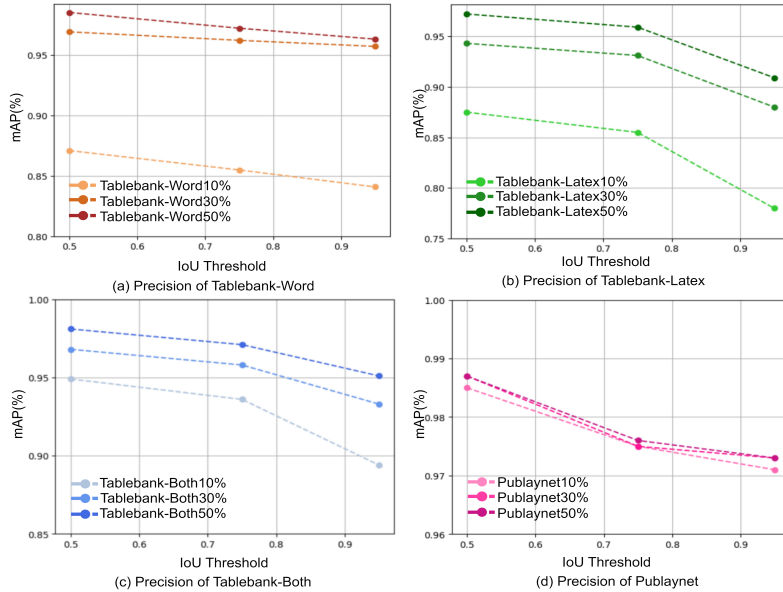
13

Fig. 6: Precision-Based Analysis of our Model's Performance. This assessment covers four distinct datasets: (a) Tablebank-Word, (b) Tablebank-Latex, (c) Tablebank-Both and (d) Publaynet. Utilizing ResNet-50 as the foundational architecture, experiments were conducted across three different data splits: 10%, 30%, and 50%.

Method	Technique	Detector	10%	30%	50%
Ren et al. [89].	Sup	Faster R-CNN	80.1	80.6	83.3
Zhu et al.[88].	Sup	Def. DETR	80.8	82.6	86.9
Sohn et al. [77]	Semi-sup	Faster R-CNN	82.4	83.8	87.1
Liu et al. [90]	Semi-sup	Faster R-CNN	83.9	86.4	88.5
Tang et al. [91]	Semi-sup	Faster R-CNN	83.4	86.2	87.9
Xu et al. [11]	Semi-sup	Faster R-CNN	83.6	86.8	89.6
Shehzadi et al.[19]	Semi-sup	Def. DETR	84.2	86.8	91.8
Our	Semi-sup	Def. DETR	89.4	93.3	95.1

Table 2: Comparison of the performance of previous supervised and semi-supervised methodologies. The ResNet-50 backbone was used to train the supervised models, Deformable-DETR(Def. DETR) and Faster R-CNN, using 10%, 30%, and 50% of the TableBank-both dataset. The same splits and datasets are utilized for semi-supervised assessment. It's important to note that all results are given in terms of mAP (0.5:0.95).

trained on 10%, 30%, and 50% labeled data. Our method stands out by implementing both one-to-many and one-to-one assignment strategies, which generate superior pseudo-labels for unlabeled data. Notably, on just 10% labeled data, our semi-supervised approach outperforms the previous supervised baseline network by 8.6 mAP points and the prior semi-supervised method by 5.2 mAP points.

5.2 PubLayNet

In this subsection, we delve into the experimental results for table detection in the PubLayNet dataset [16], exploring the impact of different amounts of labeled data. We also compare our transformer-based semi-supervised method with previous supervised and semi-supervised methods, offering a detailed evaluation of their performance.

Dataset	Label-percent	mAP	AP^{50}	AP^{75}	AR_L
	10%	97.4	98.5	98.3	97.6
publaynet	30%	97.9	98.7	98.5	98.1
	50%	98.6	98.9	99.3	99.2

Table 3: Evaluation results for the table class in the PubLayNet dataset using three different data splits: 10%, 30%, and 50%.

The results of our semi-supervised approach for table detection in PubLayNet are presented in Table 3, with performance measured across various quantities of labeled data. Notably, the mAP values achieved are 97.4%, 97.9%, and 98.6% for our semi-supervised approach trained with 10%, 30%, and 50% labeled data, respectively.

Method	mAP			AR_L		
	10%	**30%**	**50%**	**10%**	**30%**	**50%**
Shehzadi et al.[19]	88.4	90.3	92.8	91.0	93.2	96.0
Ours	97.4	97.9	98.6	97.6	98.1	99.2

Table 4: Comparison of our approach with previous table detection semi-supervised approach on PubLayNet data using the splits of 10%, 30%, and 50% of labeled data.

In Table 4, we compare our semi-supervised approach with previous semi supervised table detection appraoch. For mAP, with 10% labeled data, the improvement is 8.9 percentage points (from 88.4% to 97.4%), at 30% it's 7.6 points (from 90.3% to 97.9%), and at 50% the improvement is 5.8 points (from 92.8% to 98.6%). For AR_L, the gains are even more significant: with 10% labeled data, the improvement is 6.6 points (from 91.0% to 97.6%), at 30% it's 5.9 points (from 93.2% to 98.1%), and at 50% it's a substantial 3.2 point increase (from 96.0% to 99.2%).

Table 5 presents a comprehensive evaluation of both supervised and semi-supervised networks with specific emphasis on the table class from the PubLayNet dataset [16]. We additionally compare a supervised deformable-DETR model, which was trained on varying proportions of labeled data (10%, 30%, and 50%) from the PubLayNet table class, with our semi-supervised approach employing the deformable transformer [88]. Notably, our methodology produces comparable outcomes, eliminating the need for post-processing steps like NMS.

Method	Technique	Detector	10%	30%	50%
Ren et al. [89]	Sup	Faster R-CNN	93.6	95.6	95.9
Zhu et al. [88]	Sup	Def. DETR	93.9	96.2	97.1
Sohn et al. [77]	Semi-sup	Faster R-CNN	95.8	96.9	97.8
Liu et al.[90]	Semi-sup	Faster R-CNN	96.1	97.4	98.1
Tang et al.[91]	Semi-sup	Faster R-CNN	96.7	97.9	98.0
Xu et al. [11]	Semi-sup	Faster R-CNN	96.5	98.1	98.5
Our	Semi-sup	Def. DETR	98.5	98.7	98.9

Table 5: Performance comparison with earlier supervised and semi-supervised approaches. Here, Deformable-DETR is represented as Def. DETR. All the supervised and unsupervised networks are trained on 10%, 30%, and 50% of PubLayNet table class data. The results are presented in terms of $AP50$ at an IoU threshold of 0.5.

5.3 ICDAR-19

We also evaluate our semi-supervised approach on ICDAR-19 [17] Modern Track A dataset. Table 6 shows the results of our approach on the ICDAR-19 dataset, utilizing 10%, 30%, and 50% of annotated data. Table 7 provides a comprehensive comparison of our approach with previous supervised and semi-supervised networks at IOU thresholds of 0.8 and 0.9.

Dataset	Label-percent	mAP	AP^{50}	AP^{75}	AR_L
icdar	10%	60.2	64.7	61.4	91.8
	30%	64.8	69.0	65.6	93.7
	50%	65.4	68.6	67.1	96.4

Table 6: Evaluation results for the ICDAR-19 dataset using three different data splits: 10%, 30%, and 50%.

These approaches are evaluated based on Recall, Precision, and F1-Score. Our approach shows a high recall for table detection, indicating its strength in identifying nearly all actual tables at both IOU thresholds of 0.8 and 0.9. However, this comes

Method	Approch	IOU = 0.8			IOU = 0.9		
		Recall	Precision	F1-Score	Recall	Precision	F1-Score
TableRadar[17]	Sup	94.0	95.0	94.5	89.0	90.0	89.5
NLPR-PAL[17]	Sup	93.0	93.0	93.0	86.0	86.0	86.0
Lenovo Ocean[17]	Sup	86.0	88.0	87.0	81.0	82.0	81.5
CDeC-Net[69]	Sup	93.4	95.3	94.4	90.4	92.2	91.3
HybridTabNet[25]	Sup	93.3	92.0	92.8	90.5	89.5	90.2
Shehzadi et al.[19]	Semi-sup	71.1	82.3	76.3	66.3	76.8	71.2
Ours	Semi-sup	95.0	61.4	74.5	93.0	60.7	73.4

Table 7: Performance comparison between the proposed semi-supervised approach (with 50% labels) and previous state-of-the-art results on the ICDAR 19 Track A dataset with IOU thresholds of 0.8 and 0.9.
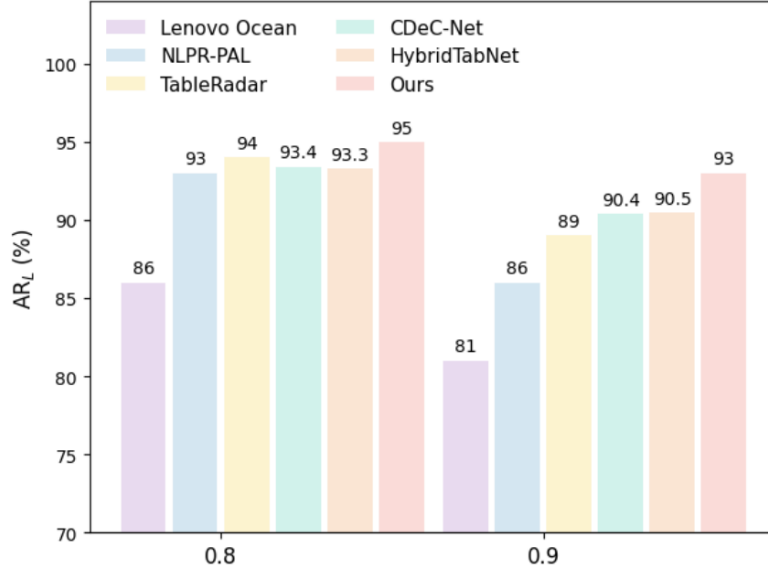
Fig. 7: Comparing our approach with other supervised approaches based on recall on the ICDAR 19 Track A dataset with IOU thresholds of 0.8 and 0.9.

with lower precision, suggesting it also picks up some false positives. Despite this, the approach effectively ensures that few actual tables are missed. Moreover, Fig. 7 shows the recall results of a comparison of our approach with the previous approach on the ICDAR-19 dataset. Here, we can observe that our approach shows the highest recall results.

Method	Technique	Labels	TableBank	PubLayNet
CDeC-Net [69]	Sup	100%	96.5	97.8
EDD [92]	Sup	100%	86.0	89.9
TableFormer [93]	Sup	100%	89.6	96.7
CasTabDetectoRS [24]	Sup	100%	95.3	-
VSR [94]	Sup	100%	-	95.9
Shehzadi et al.[19]	Semi-sup	30%	86.8	90.3
Our	Semi-sup	30%	93.3	97.9

Table 8: Performance comparison with previous supervised and semi-supervised table detection methods on TableBank-both and PubLaynet data. The results are the mean average precision (mAP) across the range from 0.5 to 0.95.

In addition, we conducted a comparative analysis of our semi-supervised method using 30% labeled data against earlier techniques in table detection and document analysis on TableBank and PubLayNet datasets (refer to Table 8). Notably, we observe

that even with only 30% labeled data, our approach yielded results comparable to those obtained by previous supervised methods.

6 Ablation Study

In this subsection, we evaluate the main components of our approach. All ablation studies utilize a ResNet-50 backbone along with a dataset subset consisting of 10% labeled images of PubLayNet data.

6.1 Impact of the quantity of learnable queries

: We examine the effect of the quantity of queries fed to the decoder. Optimal performance is achieved when employing 30 queries in a one-to-one and 400 queries in a one-to-many assignment strategy. Deviating from these values leads to a decline in performance, as observed in Table 9. We selected the 30 queries in one-to-one assignment as in the baseline; it also give the best results in our case. We then vary the queries in one-to-many assignment strategy and get the best results at 400. Deviating from this results in a performance decrease due to overfitting and overlapping predictions.

Method	# query	mAP	AP^{50}	AP^{75}	AR_L
Shehzadi et al. [19]	o2o=30,o2m=0	88.4	98.5	97.3	91.0
	o2o=30,o2m=200	89.8	95.2	94.2	93.4
Ours	**o2o=30,o2m=400**	**97.4**	**98.5**	**98.3**	**97.6**
	o2o=30,o2m=600	90.9	97.2	95.9	95.9

Table 9: Comparison of our approach with semi-supervised deformable DETR (Def. DETR) based on varying numbers of queries as a hyperparameter.

6.2 Pseudo-labeling confidence threshold

: The filtering threshold is a critical parameter influencing the quality and quantity of pseudo-labels filtered out during semi-supervised learning. Increasing the threshold reduces the number of pseudo-labels fed to the student module, but those are of higher quality. Conversely, lowering the threshold leads to more pseudo-labels filtered out, which also increases noisy pseudo-labels. Table 10 presents a comprehensive summary of the effects observed across several threshold values, from 0.5 to 0.8. According to the results, the most suitable threshold value is 0.7.

6.3 Impact of one-to-one and one-to-many matching strategies

In one-to-one matching, each ground truth matches the most similar prediction. On the other hand, one-to-many matching, as shown in Fig. 8, allows multiple predictions to match with a single ground truth, enhancing training but potentially causing overlapping predictions during testing. To balance these factors, our strategy involves using one-to-one and one-to-many matching during training and exclusively one-to-one matching during testing to ensure accurate predictions without the NMS step. Table 11

T	mAP	AP^{50}	AP^{75}
0.5	96.9	98.3	97.2
0.6	96.2	97.6	96.0
0.7	**97.4**	**98.5**	**98.3**
0.8	96.8	98.3	97.2

Table 10: Analysis of pseudo-labeling confidence thresholds and their performance impact, highlighting top-performing results in bold.

Method	mAP	AP^{50}	AP^{75}	AR$_{L}$
H-DETR [95]	95.2	97.7	96.8	98.5
Co-DETR[96]	96.3	98.2	97.6	98.9
Ours	**97.4**	**98.5**	**98.3**	**97.6**

Table 11: Comparison of our approach with other models using PubLayNet, employing a one-to-many assignment strategy to enhance object detection performance without the use of NMS. Here, highest values are shown in bold.
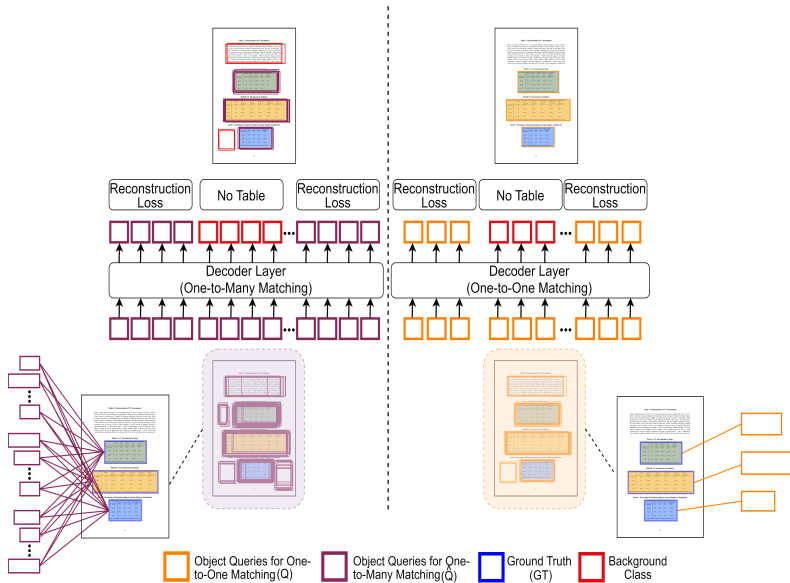

Fig. 8: Performance analysis conducted by inputting 30 one-to-one queries and 400 one-to-many queries into the decoder layer. Purple rectangles represent one-to-many queries, orange rectangles denote one-to-one queries, blue rectangles indicate ground-truth and red rectangles signify the background class.

compares our semi-supervised one-to-many matching strategy with previous supervised one-to-many matching strategies. In Table 12, We also observe the effect of one-to-one

o2m	o2o	NMS	Training Time(h)	FPS	mAP	AP^{50}	AP^{75}	AR$_{L}$
✗	✓	✗	8.74	4.36	96.0	97.8	96.8	98.9
✓	✗	✓	11.33	4.36	95.5	96.7	96.5	97.5
✓	✓	✗	**10.99**	**4.36**	**96.1**	**97.9**	**96.9**	**99.0**

Table 12: Comparative analysis of one-to-many(o2m) and one-to-one(o2o) matching strategies. It includes NMS-based matching, training time in hours, frames per second (FPS), The results are with highest values in bold without using augmented GT in one-to-many matching strategy.

and one-to-many matching strategies. One-to-one matching takes less training time as

it doesn't need an NMS step but shows low performance. One-to-many matching needs longer training time as an NMS step is needed to remove duplicate predictions. We first employ one-to-many matching to improve performance by selecting high-quality pseudo-labels and then shift to one-to-one matching to remove duplicate predictions. Our approach does not need NMS and improves performance.

7 Conclusion

In conclusion, our study addresses the challenges inherent in DETR-based object detectors within semi-supervised object detection. We recognized the limitations of the conventional one-to-one assignment approach, particularly its inefficiency in handling imprecise pseudo-labels. To overcome these issues, we introduced a semi-supervised, transformer-based approach for semi-supervised object detection that operates end-to-end. Our method integrates one-to-one and one-to-many matching strategies, improving overall performance. We generate the pseudo-labels by employing two distinctive modules, student and teacher. Through the iterative update process facilitated by an Exponential Moving-Average (EMA) function, these modules enhanced one another, resulting in accurate classification and bounding box predictions. The experimental findings shows the effectiveness of our approach, surpassing the performance of supervised and semi-supervised approaches across various labeling ratios on TableBank and the PubLayNet training data. In future, we aim to extend the semi-supervised learning framework based on transformers to tackle the task of table structure recognition. This research represents significant progress in advancing more robust and efficient methodologies for semi-supervised object detection within document analysis tasks.

Acknowledgements

References

[1] Mori, S., Suen, C.Y., Yamamoto, K.: Historical review of ocr research and development. Proceedings of the IEEE **80**(7), 1029–1058 (1992)

[2] Nguyen, T.T.H., Jatowt, A., Coustaty, M., Doucet, A.: Survey of post-ocr processing approaches. ACM Computing Surveys (CSUR) **54**(6), 1–37 (2021)

[3] Singh, A., Bacchuwar, K., Bhasin, A.: A survey of ocr applications. International Journal of Machine Learning and Computing **2**(3), 314 (2012)

[4] Bourbakis, N., Burton, M.: Ocr techniques for the design of non-conventional typed text reading systems. Engineering Applications of Artificial Intelligence **4**(3), 191–204 (1991)

[5] Bhatt, P.P., Patel, I.: Optical character recognition using deep learning-a technical review. National Journal of System and Information Technology **11**(1), 55 (2018)

[6] Prasad, D., Gadpal, A., Kapadni, K., Visave, M., Sultanpure, K.: Cascadetabnet: An approach for end to end table detection and structure recognition from image-based documents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 572–573 (2020)

[7] Shekar, K.C., Cross, M.A., Vasudevan, V.: Optical character recognition and neural machine translation using deep learning techniques. In: Innovations in Computer Science and Engineering: Proceedings of 8th ICICSE, pp. 277–283 (2021). Springer

[8] Shehzadi, T., Hashmi, K.A., Stricker, D., Afzal, M.Z.: Object Detection with Transformers: A Review (2023)

[9] Xue, C., Zhang, W., Hao, Y., Lu, S., Torr, P.H., Bai, S.: Language matters: A weakly supervised vision-language pre-training approach for scene text detection and spotting. In: European Conference on Computer Vision, pp. 284–302 (2022). Springer

[10] Amrhein, C., Clematide, S.: Supervised ocr error detection and correction using statistical and neural machine translation methods. Journal for Language Technology and Computational Linguistics (JLCL) **33**(1), 49–76 (2018)

[11] Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., Liu, Z.: End-to-end semi-supervised object detection with soft teacher. CoRR **abs/2106.09018** (2021) 2106.09018

[12] Hosang, J., Benenson, R., Schiele, B.: Learning non-maximum suppression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4507–4515 (2017)

[13] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision, pp. 213–229 (2020). Springer

[14] Fang, R., Gao, P., Zhou, A., Cai, Y., Liu, S., Dai, J., Li, H.: Feataug-detr: Enriching one-to-many matching for detrs with feature augmentation. arXiv preprint arXiv:2303.01503 (2023)

[15] Li, S., Li, M., Li, R., He, C., Zhang, L.: One-to-few label assignment for end-to-end dense detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7350–7359 (2023)

[16] Zhong, X., Tang, J., Yepes, A.J.: Publaynet: largest dataset ever for document layout analysis. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1015–1022 (2019). https://doi.org/10.1109/ICDAR.

2019.00166 . IEEE

[17] Gao, L., Huang, Y., Déjean, H., Meunier, J.-L., Yan, Q., Fang, Y., Kleber, F., Lang, E.: Icdar 2019 competition on table detection and recognition (ctdar). In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1510–1515 (2019). IEEE

[18] Li, M., Cui, L., Huang, S., Wei, F., Zhou, M., Li, Z.: Tablebank: Table benchmark for image-based table detection and recognition. CoRR **abs/1903.01949** (2019) 1903.01949

[19] Shehzadi, T., Azeem Hashmi, K., Stricker, D., Liwicki, M., Zeshan Afzal, M.: Towards end-to-end semi-supervised table detection with deformable transformer. In: Fink, G.A., Jain, R., Kise, K., Zanibbi, R. (eds.) Document Analysis and Recognition - ICDAR 2023, pp. 51–76. Springer, Cham (2023)

[20] Tupaj, S., Shi, Z., Chang, C.H., Alam, H.: Extracting tabular information from text files. EECS Department, Tufts University, Medford, USA **1** (1996)

[21] Hirayama, Y.: A method for table structure analysis using dp matching. In: Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 2, pp. 583–5862 (1995). https://doi.org/10.1109/ICDAR.1995.601964

[22] Kieninger, T.G.: Table structure recognition based on robust block segmentation. In: Document Recognition V, vol. 3305, pp. 22–32 (1998). SPIE

[23] Schreiber, S., Agne, S., Wolf, I., Dengel, A., Ahmed, S.: Deepdesrt: Deep learning for detection and structure recognition of tables in document images. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 01, pp. 1162–1167 (2017). https://doi.org/10.1109/ICDAR.2017.192

[24] Hashmi, K.A., Pagani, A., Liwicki, M., Stricker, D., Afzal, M.Z.: Castabdetectors: Cascade network for table detection in document images with recursive feature pyramid and switchable atrous convolution. Journal of Imaging **7** (2021)

[25] Nazir, D., Hashmi, K.A., Pagani, A., Liwicki, M., Stricker, D., Afzal, M.Z.: Hybridtabnet: Towards better table detection in scanned document images. Applied Sciences **11**(18) (2021) https://doi.org/10.3390/app11188396

[26] Itonori, K.: Table structure recognition based on textblock arrangement and ruled line position. In: Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR '93), pp. 765–768 (1993). https://doi.org/10.1109/ICDAR.1993.395625

[27] Chandran, S., Kasturi, R.: Structural recognition of tabulated data. In: Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR '93), pp. 516–519 (1993). https://doi.org/10.1109/ICDAR.1993.395683

[28] Pyreddy, P., Croft, W.B.: Tintin: a system for retrieval in text tables. In: Digital Library (1997)

[29] Pivk, A., Cimiano, P., Sure, Y., Gams, M., Rajkovič, V., Studer, R.: Transforming arbitrary tables into logical form with tartar. Data & Knowledge Engineering **60**(3), 567–595 (2007) https://doi.org/10.1016/j.datak.2006.04.002

[30] Hu, J., Kashi, R.S., Lopresti, D.P., Wilfong, G.: Medium-independent table detection. In: Document Recognition and Retrieval VII, vol. 3967, pp. 291–302 (1999). SPIE

[31] Coüasnon, B., Lemaitre, A.: Recognition of tables and forms. In: Handbook of Document Image Processing and Recognition (2014). https://api.semanticscholar.org/CorpusID:31528017

[32] Jorge, A.M., Torgo, L., *et al.*: Design of an end-to-end method to extract information from tables. International Journal of Document Analysis and Recognition (IJDAR) **8**(2), 144–171 (2006)

[33] Embley, D.W., Hurst, M., Lopresti, D., Nagy, G.: Table-processing paradigms: a research survey. International Journal of Document Analysis and Recognition (IJDAR) **8**(2), 66–86 (2006)

[34] Cesarini, F., Marinai, S., Sarti, L., Soda, G.: Trainable table location in document images. In: 2002 International Conference on Pattern Recognition, vol. 3, pp. 236–2403 (2002). https://doi.org/10.1109/ICPR.2002.1047838

[35] Silva, A.C.e.: Learning rich hidden markov models in document analysis: Table location. In: 2009 10th International Conference on Document Analysis and Recognition, pp. 843–847 (2009). https://doi.org/10.1109/ICDAR.2009.185

[36] Silva, A.: Parts that add up to a whole: a framework for the analysis of tables. Edinburgh University, UK (2010)

[37] Kasar, T., Barlas, P., Adam, S., Chatelain, C., Paquet, T.: Learning to detect tables in scanned document images using line information. In: 2013 12th International Conference on Document Analysis and Recognition, pp. 1185–1189 (2013). IEEE

[38] Shehzadi, T., Majid, A., Hameed, M., Farooq, A., Yousaf, A.: Intelligent predictor using cancer-related biologically information extraction from cancer transcriptomes. In: 2020 International Symposium on Recent Advances in Electrical Engineering & Computer Sciences (RAEE & CS), vol. 5, pp. 1–5 (2020)

[39] Yousaf, A., Shehzadi, T., Farooq, A., Ilyas, K.: Protein active site prediction for early drug discovery and designing. International Review of Applied Sciences and Engineering **13**(1), 98–105 (2021)

[40] Saeed, W., Saleh, M.S., Gull, M.N., Raza, H., Saeed, R., Shehzadi, T.: Geometric features and traffic dynamic analysis on 4-leg intersections. International Review of Applied Sciences and Engineering (2023)

[41] Minouei, M., Hashmi, K.A., Soheili, M.R., Afzal, M.Z., Stricker, D.: Continual learning for table detection in document images. Applied Sciences **12**(18) (2022) https://doi.org/10.3390/app12188969

[42] Kölsch, A., Afzal, M.Z., Ebbecke, M., Liwicki, M.: Real-time document image classification using deep cnn and extreme learning machines. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 01, pp. 1318–1323 (2017). https://doi.org/10.1109/ICDAR.2017.217

[43] Sinha, S., Hashmi, K.A., Pagani, A., Liwicki, M., Stricker, D., Afzal, M.Z.: Rethinking learnable proposals for graphical object detection in scanned document images. Applied Sciences **12**(20) (2022) https://doi.org/10.3390/app122010578

[44] Naik, S., Hashmi, K.A., Pagani, A., Liwicki, M., Stricker, D., Afzal, M.Z.: Investigating attention mechanism for page object detection in document images. Applied Sciences **12**(15) (2022) https://doi.org/10.3390/app12157486

[45] Hashmi, K.A., Pagani, A., Liwicki, M., Stricker, D., Afzal, M.Z.: Cascade network with deformable composite backbone for formula detection in scanned document images. Applied Sciences **11**(16) (2021) https://doi.org/10.3390/app11167610

[46] Yang, X., Yümer, M.E., Asente, P., Kraley, M., Kifer, D., Giles, C.L.: Learning to extract semantic structure from documents using multimodal fully convolutional neural network. CoRR **abs/1706.02337** (2017) 1706.02337

[47] He, D., Cohen, S., Price, B., Kifer, D., Giles, C.L.: Multi-scale multi-task fcn for semantic page segmentation and table detection. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 01, pp. 254–261 (2017). https://doi.org/10.1109/ICDAR.2017.50

[48] Kavasidis, I., Palazzo, S., Spampinato, C., Pino, C., Giordano, D., Giuffrida, D., Messina, P.: A saliency-based convolutional neural network for table and chart detection in digitized documents. CoRR **abs/1804.06236** (2018) 1804.06236

[49] Paliwal, S., D, V., Rahul, R., Sharma, M., Vig, L.: Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images. CoRR **abs/2001.01469** (2020) 2001.01469

[50] Smock, B., Pesala, R., Abraham, R.: PubTables-1M: Towards comprehensive table extraction from unstructured documents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4634–4642 (2022)

[51] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. CoRR **abs/1411.4038** (2014) 1411.4038

[52] Li, X.-H., Yin, F., Liu, C.-L.: Page object detection from pdf document images by deep structured prediction and supervised clustering. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 3627–3632 (2018). https://doi.org/10.1109/ICPR.2018.8546073

[53] Holecek, M., Hoskovec, A., Baudis, P., Klinger, P.: Line-items and table understanding in structured documents. CoRR **abs/1904.12577** (2019) 1904.12577

[54] Yi, X., Gao, L., Liao, Y., Zhang, X., Liu, R., Jiang, Z.: Cnn based page object detection in document images. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 01, pp. 230–235 (2017). https://doi.org/10.1109/ICDAR.2017.46

[55] Hao, L., Gao, L., Yi, X., Tang, Z.: A table detection method for pdf documents based on convolutional neural networks. 2016 12th IAPR Workshop on Document Analysis Systems (DAS), 287–292 (2016)

[56] Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. CoRR **abs/1708.02002** (2017) 1708.02002

[57] Fang, Y., Liao, B., Wang, X., Fang, J., Qi, J., Wu, R., Niu, J., Liu, W.: You only look at one sequence: Rethinking transformer in vision through object detection. CoRR **abs/2106.00666** (2021) 2106.00666

[58] Girshick, R.B.: Fast R-CNN. CoRR **abs/1504.08083** (2015) 1504.08083

[59] Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. CoRR **abs/1506.01497** (2015) 1506.01497

[60] He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. CoRR **abs/1703.06870** (2017) 1703.06870

[61] Cai, Z., Vasconcelos, N.: Cascade R-CNN: delving into high quality object detection. CoRR **abs/1712.00726** (2017) 1712.00726

[62] Shehzadi, T., Stricker, D., Afzal, M.Z.: A Hybrid Approach for Document Layout Analysis in Document images (2024)

[63] Shehzadi, T., Sarode, S., Stricker, D., Afzal, M.Z.: Towards End-to-End Semi-Supervised Table Detection with Semantic Aligned Matching Transformer (2024)

[64] Kanchi, S., Pagani, A., Mokayed, H., Liwicki, M., Stricker, D., Afzal, M.Z.: Emmdocclassifier: Efficient multimodal document image classifier for scarce data. Applied Sciences **12**(3) (2022) https://doi.org/10.3390/app12031457

[65] Gilani, A., Qasim, S.R., Malik, I., Shafait, F.: Table detection using deep learning. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 01, pp. 771–776 (2017). https://doi.org/10.1109/ICDAR.2017.131

[66] Prasad, D., Gadpal, A., Kapadni, K., Visave, M., Sultanpure, K.: Cascadetabnet: An approach for end to end table detection and structure recognition from image-based documents. CoRR **abs/2004.12629** (2020) 2004.12629

[67] Arif, S., Shafait, F.: Table detection in document images using foreground and background features. In: 2018 Digital Image Computing: Techniques and Applications (DICTA), pp. 1–8 (2018). https://doi.org/10.1109/DICTA.2018.8615795

[68] Siddiqui, S.A., Malik, M.I., Agne, S., Dengel, A., Ahmed, S.: Decnt: Deep deformable cnn for table detection. IEEE Access **6**, 74151–74161 (2018) https://doi.org/10.1109/ACCESS.2018.2880211

[69] Agarwal, M., Mondal, A., Jawahar, C.V.: Cdec-net: Composite deformable cascade network for table detection in document images. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 9491–9498 (2021). https://doi.org/10.1109/ICPR48806.2021.9411922

[70] Shehzadi, T., Hashmi, K.A., Stricker, D., Liwicki, M., Afzal, M.Z.: Bridging the performance gap between detr and r-cnn for graphical object detection in document images. arXiv preprint arXiv:2306.13526 (2023)

[71] Jeong, J., Lee, S., Kim, J., Kwak, N.: Consistency-based semi-supervised learning for object detection. Advances in neural information processing systems **32** (2019)

[72] Tang, P., Ramaiah, C., Xu, R., Xiong, C.: Proposal learning for semi-supervised object detection. CoRR **abs/2001.05086** (2020) 2001.05086

[73] Radosavovic, I., Dollár, P., Girshick, R.B., Gkioxari, G., He, K.: Data distillation: Towards omni-supervised learning. CoRR **abs/1712.04440** (2017) 1712.04440

[74] Zoph, B., Ghiasi, G., Lin, T.-Y., Cui, Y., Liu, H., Cubuk, E.D., Le, Q.: Rethinking pre-training and self-training. Advances in neural information processing systems **33**, 3833–3845 (2020)

[75] Shehzadi, T., Hashmi, K.A., Pagani, A., Liwicki, M., Stricker, D., Afzal, M.Z.: Mask-aware semi-supervised object detection in floor plans. Applied Sciences **12**(19) (2022)

[76] Shehzadi, T., Hashmi, K.A., Stricker, D., Afzal, M.Z.: Sparse semi-detr: Sparse learnable queries for semi-supervised object detection. arXiv preprint arXiv:2404.01819 (2024)

[77] Sohn, K., Zhang, Z., Li, C., Zhang, H., Lee, C., Pfister, T.: A simple semi-supervised learning framework for object detection. CoRR **abs/2005.04757** (2020) 2005.04757

[78] Li, Y., Huang, D., Qin, D., Wang, L., Gong, B.: Improving object detection with selective self-supervised self-training. CoRR **abs/2007.09162** (2020) 2007.09162

[79] Wei, X., Yan, S., Wang, X., Guizani, M., Du, X.: Stac: A spatio-temporal approximate method in data collection applications. Pervasive and Mobile Computing **73**, 101371 (2021)

[80] Liu, Y., Ma, C., He, Z., Kuo, C., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. CoRR **abs/2102.09480** (2021) 2102.09480

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

[82] Zhou, Q., Yu, C., Wang, Z., Qian, Q., Li, H.: Instant-teaching: An end-to-end semi-supervised object detection framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4081–4090 (2021)

[83] Wang, P., Cai, Z., Yang, H., Swaminathan, G., Vasconcelos, N., Schiele, B., Soatto, S.: Omni-detr: Omni-supervised object detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9367–9376 (2022)

[84] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[85] Zhao, J., Ghosh, S., Bharadwaj, A., Ma, C.-Y.: When does the student surpass the teacher? federated semi-supervised learning with teacher-student ema. arXiv preprint arXiv:2301.10114 (2023)

[86] Szegedy, C., Ioffe, S., Vanhoucke, V.: Inception-v4, inception-resnet and the impact of residual connections on learning. CoRR **abs/1602.07261** (2016) 1602.07261

[87] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)

[88] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. CoRR **abs/2010.04159** (2020) 2010.04159

[89] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)

[90] Liu, Y.-C., Ma, C.-Y., He, Z., Kuo, C.-W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. arXiv preprint arXiv:2102.09480 (2021)

[91] Tang, Y., Chen, W., Luo, Y., Zhang, Y.: Humble teachers teach better students for semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3132–3141 (2021)

[92] Zhong, X., ShafieiBavani, E., Jimeno Yepes, A.: Image-based table recognition: data, model, and evaluation. In: European Conference on Computer Vision, pp. 564–580 (2020). Springer

[93] Nassar, A., Livathinos, N., Lysak, M., Staar, P.: Tableformer: Table structure understanding with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4614–4623 (2022)

[94] Zhang, P., Li, C., Qiao, L., Cheng, Z., Pu, S., Niu, Y., Wu, F.: VSR: A unified framework for document layout analysis combining vision, semantics and relations. CoRR **abs/2105.06220** (2021) 2105.06220

[95] Jia, D., Yuan, Y., He, H., Wu, X., Yu, H., Lin, W., Sun, L., Zhang, C., Hu, H.: Detrs with hybrid matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19702–19712 (2023)

[96] Zong, Z., Song, G., Liu, Y.: Detrs with collaborative hybrid assignments training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6748–6758 (2023)