DocSemi: Efficient Document Layout Analysis with Guided Queries

Tahira Shehzadi[1,2,3] Ifza Ifza[1,3] 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 Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany
tahira.shehzadi@dfki.de

Abstract

Recent advancements in deep learning have significantly enhanced object detection, especially in semi-supervised contexts. We propose DocSemi, a novel end-to-end semi-supervised document layout analysis framework that leverages focused attention for efficiency, alongside a hybrid query mechanism and guided queries for accurate document detection. We conduct extensive experiments on several benchmark datasets, including PubLayNet, PubTables, and TableBank covering a wide range of document structures. Notably, our approach requires significantly fewer training epochs and uses adaptive query selection for better generalization. Experimental results on the PubLayNet dataset under 10% labeled data settings show that our model achieves competitive performance, with a mean Average Precision (mAP) of up to 96.2% and an Average Recall of 98.3%, outperforming prior methods in efficiency while maintaining high accuracy. These results highlight the effectiveness of integrating focused attention and guided queries for accurate and efficient semi-supervised document layout analysis.

1. Introduction

Document layout analysis (DLA) [5, 6, 67] focuses on identifying and organizing key structural elements such as headers, paragraphs, images, and tables within a document to ensure a well-structured format. This task is critical for applications such as document digitization [11] and content extraction [12], where understanding the layout enables effective processing. However, accurately detecting and classifying components across diverse document formats [3, 29] is challenging due to significant variations in layouts based on the source and content. The complexity and diversity of document structures underscore the need for robust DLA methods [6] that generalize well across document types while maintaining precision. However, a major challenge lies in the limited availability of labeled data, as manually annotating large-scale datasets [37] is time consuming and labor intensive. Semi-supervised learning [4] addresses this by using both labeled and unlabeled data, improving the generalizability and efficiency of the model in varied layouts.

Early approaches to DLA relied heavily on rule-based methods [15, 17, 77] that used manually defined criteria to identify structural elements. While effective for consistently formatted documents, these methods struggled with borderless tables and complex layouts. OCR systems [68, 80], built on rule-based frameworks, focused mainly on text extraction but lacked accuracy in detecting and organizing non-text elements like images and tables. Deep learning methods [2, 42], especially Convolutional Neural Networks (CNNs) [19, 34, 76], emerged to learn layout patterns directly from data, reducing reliance on manual rules. For example, Faster R-CNN [44] uses a two-stage process to generate and classify region proposals, improving detection across diverse layouts. However, these approaches require extensive labeled data, which is often impractical to obtain.

Transformer-based models [24, 38] have revolutionized DLA by leveraging self-attention mechanisms [86] to capture complex relationships within documents. Unlike CNNs that focus on local information, transformers process entire documents, better recognizing intricate layouts and spatial relationships. Models such as Vision Transformer (ViT) [27] and LayoutLM [21] combine visual and textual data, enhancing detection and classification. Pre-training on large datasets further improves their generalization across document types, increasing accuracy and robustness.

In object detection, DETR-based frameworks [71] utilize one-to-one assignment techniques [30] via bipartite matching [7] to match ground-truth bounding boxes with candidate proposals efficiently. While effective with accurate ground-truth data, this approach faces challenges in semi-supervised settings due to training inefficiencies. One-to-many assignment strategies [18] can improve proposal quality but risk producing duplicate predictions.

To overcome these issues, we propose a DETR-based semi-supervised object detection (SSOD) framework [63, 79] integrating both assignment methods through a Hybrid

Matching approach [58]. The one-to-one assignment uses 30 fixed object queries for stable matching, while the one-to-many assignment employs adaptive queries ranging from 3 to 9 for flexible detection and localization. This dynamic query configuration [64] enhances training efficiency and enables the generation of high-quality pseudo-labels [41] for iterative refinement.

To improve the model's ability to capture salient features while minimizing noise, we incorporate a Focused Attention Network (FAN) mechanism [10, 83]. FAN enables the model to concentrate on relevant structural elements within complex document layouts, thereby enhancing detection accuracy in document layout analysis (DLA). FAN is integrated into both the student and teacher encoder to increase efficiency. In parallel, we adopt the guided queries technique [78] to suppress noise—such as artifacts from low-quality scans—and steer object queries toward stable and informative regions. Guided queries also enhance the quality of pseudo-labels by stabilizing query guidance. Together, these components form a robust and scalable semi-supervised framework, enabling both models [1, 48] to produce higher-quality predictions and achieve improved overall accuracy

To sum up, this study provides significant advances in the following essential areas:

- We introduce a novel end-to-end semi-supervised approach for document layout analysis, leveraging focused attention and guided queries to enhance performance with limited labeled data.
- Our approach employs adaptive hybrid matching strategy, which combines both one-on-one and one-to-many assignment methods with adaptive refinement, optimizing the model for both labeled and unlabeled data scenarios.
- A comprehensive evaluation using diverse datasets, including PubLayNet, PubTables, and TableBank demonstrates the effectiveness of our approach, achieving competitive results compared to state-of-the-art approaches.

2. Related Work

The study of document analysis has advanced significantly over the years, with various methods developed to improve how we extract and understand information from documents. Initially, document analysis depended on rule-based systems that used fixed rules to identify and process document elements [23, 26]. As technology progressed, machine learning techniques introduced more flexible and effective ways to handle different types of documents. The introduction of deep learning further advanced the field, providing powerful tools for tasks such as text recognition and layout analysis [36, 50]. This section summarizes the important advances in document analysis.

2.1. Heuristics Approaches to Deep Learning

In document layout analysis, the shift from heuristics to deep learning represents a significant advancement in the methods used to extract and analyze document structures. Heuristic approaches, which rely on rule-based systems and manual feature engineering [13, 15, 17, 46, 51], provided practical solutions for segmenting and classifying document elements. However, they struggled with the variety and complexity of real-world documents, requiring extensive manual setup and expert knowledge to set up and adjust the rules, making them less adaptable to diverse types of documents.

Deep learning [2, 42, 50] took over rule-based algorithms for document layout analysis, with Convolutional Neural Networks (CNNs) [19, 25, 34] becoming the main approach for segmenting document elements. For instance, Faster-RCNN [44] provided strong capabilities for detecting objects within documents and was useful for page segmentation tasks. Mask-RCNN [20] set new standards for layout segmentation, particularly for newspaper elements. Another approach, RetinaNet [33], focused on detecting keywords in document images, although it was mainly geared towards text regions. DeepDeSRT [49] introduced a breakthrough in table detection by using a unique image transformation technique to recognize table structures with a fully convolutional network. Similarly, Oliviera et al. [9] used an CNN-based framework for accurate pixel-wise segmentation of historical documents, surpassing earlier methods that used convolutional autoencoders. The ICDAR2017 POD [47] benchmark achieved top results using Faster-RCNN [44] for detecting various document elements like equations and tables, while a new cross-domain benchmark addressed the issue of domain shifts through adaptation strategies.Recently, Yang et al. set a new benchmark with a vision-based layout detection model that utilized synthetic PDF documents and a recurrent CNN with the VoVNet-v2 backbone [75], enhancing segmentation for scientific documents. Meanwhile, Li et al. [31] developed a multimodal framework for understanding structured text in documents, though it struggled with text containing similar contextual information.

2.2. State-of-the-Art Methods in DLA

Recent progress in document analysis have been greatly influenced by cutting-edge deep learning techniques, transforming the way documents are processed and understood. Transformer-based models [38, 54], such as DiT [27] and LayoutLMv3 [21], have set new standards in document analysis by leveraging sophisticated self-attention mechanisms [69] and positional embeddings [74]. DiT [27] has excelled in tasks such as document image classification and layout analysis, showing exceptional performance with large, unlabeled datasets through self-supervised learning. However, this approach may be less effective with smaller datasets, such as those from specialized sources like

PRIMA [35]. To address these limitations, the TILT [40] mechanism has been introduced, enabling simultaneous learning of textual, visual, and layout features through an encoder-decoder Transformer architecture [69]. Additionally, LayoutLMv3 [21] has achieved significant results by integrating text, layout, and visual information, though it faces challenges with smaller datasets. Other contemporary models have also employed joint pretraining to enhance various document analysis tasks, but they often struggle with domain adaptation and limited class information. The Vision Grid Transformer (VGT) [14] further advances the field with its two-stage visual pre-training process, though it comes with longer inference times.

Graph-Based Methods [55, 70] leverage the inherent structure of documents by representing them as graphs, where nodes correspond to elements such as text blocks or tables, and edges capture their relationships and spatial arrangements. This method allows a deeper understanding of the document's structure and relationships between its components. For example, Graph Neural Networks (GNNs) [84] can be employed to analyze and extract hierarchical and relational information from documents, facilitating improved layout analysis and information retrieval. Graph-based methods are particularly effective in scenarios where the document's structure is crucial, such as in complex forms or documents with Complex layouts.

Semi-supervised learning [4, 56, 60, 79] leverages both labeled and unlabeled data to improve model performance, addressing the challenge of limited labeled data in many domains. In document analysis, this approach typically employs techniques such as pseudo-labeling [41, 52, 63], where the model generates labels for unlabeled data based on its predictions, thereby expanding the training set with additional labeled examples. Consistency regularization [22] enhances robustness by encouraging the model to produce consistent predictions for augmented versions of the same input. Additionally, self-training [86] iteratively refines the model using its own predictions as pseudo-labels. These techniques collectively improve the model's ability to generalize across diverse document types and boost accuracy in tasks such as text recognition and layout analysis.

However, existing semi-supervised methods [4, 56, 59, 79] often struggle with noisy pseudo-labels and complex document layouts. To address these challenges, our approach employs a teacher-student framework [1] enhanced with Focused Attention [10, 83] to better capture layout features. We further introduce a Guided Queries strategy [78] that combines fixed and adaptive queries for improved detection. Moreover, cross-view query consistency and variance filtering techniques are incorporated to ensure that only reliable pseudo-labels guide the training process.

3. Our approach

3.1. Preliminary

Our objective is to enhance the performance of semi-supervised object detection. This scenario involves training with both labeled image sets $D_s = \{x_i^s\}_{i=1}^{N_s}$ and unlabeled image sets $D_u = \{x_i^u\}_{i=1}^{N_u}$, where N_s represents the number of labeled images and N_u represents the number of unlabeled images. For the labeled images x^s, the associated annotations y^s include information about the coordinates and categories of all bounding boxes. Our framework, illustrated in Figure 1, is built on a teacher-student architecture designed specifically for semi-supervised document object detection. The teacher and student networks [1] differ in the types of augmentations [43] applied to input data. The teacher receives weakly augmented images [81] to preserve the original layout and structural integrity of the document, while the student is trained on strongly augmented images [73]. This design enables the teacher to generate stable and reliable pseudo-labels [41, 52, 63], while encouraging the student to learn more robust and generalizable representations under diverse input conditions. Both the teacher and student encoders utilize the Focused Attention mechanism, which integrates spatial and channel-wise attention to effectively capture key layout features, as detailed in Section 3.2. Furthermore, as described in Section 3.3, a Guided Queries strategy and cross-view query consistency is applied during decoding to improve robustness against noisy predictions from unlabeled data. Section 3.4 details the loss functions that guide learning in both supervised and unsupervised settings.

3.2. Focused Attention Mechanism

Focused Attention [10, 83] is a hybrid attention mechanism that combines both multi-head self-attention [69] and channel attention [8, 72] to enhance the model's ability to focus on critical aspects of the input data.

Our approach integrates the focused Attention module within both the student and teacher encoders. As illustrated in Figure 1, a 3×3 convolution $\mathcal{C}$ with padding is first applied to the input feature map $X \in \mathbb{R}^{H\times W\times C}$ to capture local spatial features while preserving spatial dimensions:

$$X' = \mathcal{C}(X) \tag{1}$$

The resulting feature map X' is then passed through a multi-head self-attention (MHSA) mechanism. The input is projected into queries Q, keys K, and values V as follows:

$$Q = X'W_Q, \quad K = X'W_K, \quad V = X'W_V \tag{2}$$

where W_Q, W_K, and W_V are learned projection matrices.

Each attention head i computes:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

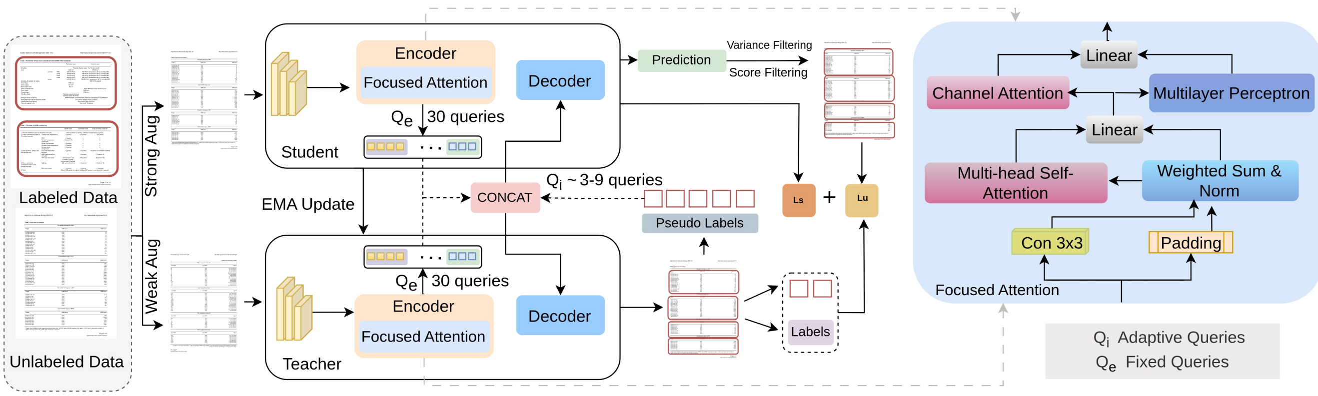

Figure 1. Our approach consists of two main modules: the student and teacher modules. The decoder in both modules utilizes a combination of one-to-one and one-to-many assignment strategies. In the teacher module, the one-to-many assignment strategy is employed to generate high-quality pseudo-labels, which are then used to enhance performance with limited labeled data. The student module, on the other hand, uses the one-to-many strategy to produce accurate predictions, while the one-to-one assignment strategy is applied to filter out redundant predictions, ensuring cleaner results. Both modules leverage focused attention mechanisms, incorporating channel attention, multi-head self-attention, and multilayer perceptrons to refine the queries (adaptive and fixed) and improve overall accuracy.

where W_i^Q, W_i^K, and W_i^V are the projection matrices specific to the i-th head.

The outputs from all heads are concatenated and projected:

$$\text{MHSA}(X') = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W_O \tag{4}$$

where W_O is a learned output projection matrix.

The output of MHSA is passed through a residual connection and normalization (e.g., LayerNorm), followed by a Multi-Layer Perceptron (MLP):

$$\hat{X} = \text{MLP}(\text{Norm}(X' + \text{MHSA}(X'))) \tag{5}$$

The MLP consists of fully connected layers interleaved with non-linear activations, further refining the representation [39, 61, 66]. To further optimize performance, channel attention [8, 72] is incorporated. The feature map is reshaped into $X_c \in \mathbb{R}^{C \times N}$, where $N = H \times W$, and channel-wise self-attention is computed as:

$$Y = \sigma \left(\frac{\text{Softmax}(Q_c)\, \text{Softmax}(K_c)^\top}{\sqrt{d}} \right) \tag{6}$$

where $Q_c, K_c \in \mathbb{R}^{d \times N}$, d is the reduced dimensionality of each channel vector, and σ denotes the sigmoid activation function. The final recalibrated output is obtained via element-wise multiplication:

$$X_{\text{out}} = Y \odot \hat{X} \tag{7}$$

where $\odot$ denotes element-wise multiplication, allowing the model to focus on the most informative feature channels.

The teacher model is updated from the student model using Exponential Moving Average (EMA) [65], which stabilizes training and produces high-quality pseudo-labels. By leveraging focused Attention in both models, we ensure continuous refinement and robust feature alignment, ultimately improving the effectiveness of document layout analysis while prioritizing essential spatial and channel-wise information throughout the learning process.

3.3. Guided Queries Strategy

We propose a hybrid query strategy that combines one-to-one and one-to-many assignment approaches to improve semi-supervised object detection. Our model uses 30 fixed queries Q_e for one-to-one matching, providing clean and precise detections, and 3–9 adaptive queries Q_i for one-to-many matching, improving coverage and recall. The encoder in both the student and teacher models generates the fixed object queries Q_e. Foreground queries from the teacher, based on high-confidence predictions from unlabeled data, serve as adaptive queries Q_i. These adaptive queries are concatenated with the fixed queries and passed through the decoder in both the teacher and student models.

This hybrid approach balances precision and flexibility, resulting in more robust and accurate layout detection:

$$Q = \text{Concat}(Q_e, Q_i) \tag{8}$$

$$\text{where} \quad Q_e \in \mathbb{R}^{30 \times d}, \quad Q_i \in \mathbb{R}^{k \times d}, \quad k \in [3, 9].$$

To mitigate the impact of noisy outputs, the Guided Queries Strategy [78] applies two filtering steps to the student's predictions on unlabeled data: *confidence score filtering* and *variance filtering*.

Confidence score filtering retains only predictions with classification confidence above a threshold τ_s:

$$\hat{y} = \{ y \mid \text{score}(y) > \tau_s \} \tag{9}$$

Method	Epochs	Queries	10% Label			30% Label			50% Label		
			mAP	AP^{50}	AP^{75}	mAP	AP^{50}	AP^{75}	mAP	AP^{50}	AP^{75}
Deformable-semi [53]	150	30	88.4	98.5	97.3	90.3	98.8	97.5	90.9	97.4	94.9
SAM-semi [57]	120	30	89.9	97.1	94.3	92.8	98.8	97.3	93.2	97.7	95.0
I. Ehsan [16]	150	30 + 400	97.4	98.5	98.3	97.9	98.7	98.5	98.6	98.9	99.3
our	12	30 + Adaptive	96.2	97.7	97.6	96.2	97.7	97.4	95.9	97.7	97.5

Table 1. Comparison of model performance on the PubLayNet table class dataset using 10%, 30%, and 50% labeled data. mAP, AP^{50}, and AP^{75} are evaluated for three models: Deformable-semi, SAM-semi, I. Ehsan, and our model.

Method	Epochs	Queries	10% Label	30% Label	50% Label
			AR_L	AR_L	AR_L
Deformable-semi [53]	150	30	91.0	93.2	96.6
SAM-semi [57]	120	30	96.6	96.0	97.3
I. Ehsan [16]	150	30 + 400	97.6	98.1	99.2
Ours	12	30 + Adaptive	98.2	98.3	98.0

Table 2. AR_L comparison across models for 10%, 30%, and 50% labeled data on PubLayNet.

Dataset	Labels	mAP	AP^{50}	AP^{75}	AR_L
	10%	96.2	97.7	97.6	98.2
PubLayNet	30%	96.2	97.7	97.4	98.3
	50%	95.9	97.7	97.5	98.0

Table 3. Performance of our model on PubLayNet across different label percentages.

Method	Approach	10%	30%	50%
Ren et al. [45]	Supervised	83.4	86.6	87.9
Zhu et al. [85]	Supervised	83.9	86.8	88.1
Ours	Semi-supervised	96.2	96.2	95.9

Table 4. Comparison of our semi-supervised method with prior approaches on the PubLayNet table class, using 10%, 30%, and 50% labeled data. Results are reported in mAP.

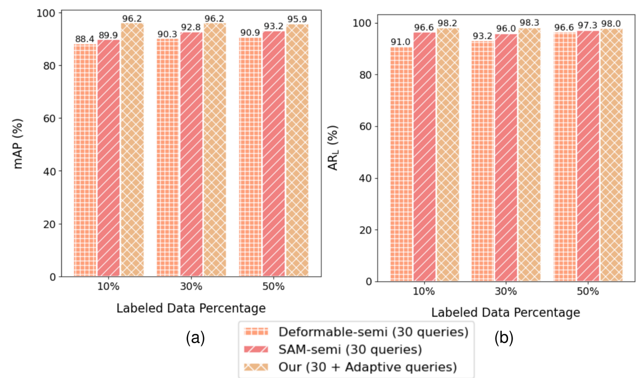

Figure 2. Performance comparison on PubLayNet for different labels (10%, 30%, and 50%) on PubLayNet, The results are presented as follows: (a) mAP comparison for Deformable-semi, SAM-semi, and Our model (b) AR_L comparison between Deformable-semi, SAM-semi, and Our model.

Variance filtering removes predictions with high uncertainty by computing the variance of predictions across multiple augmented views and discarding those with variance exceeding a threshold τ_v:

$$\mathrm{Var}(y) = \frac{1}{M}\sum_{m=1}^{M} \|y_m - \bar{y}\|_2^2 \tag{10}$$

where M is the number of augmented views, y_m is the prediction from the m-th view, and

$$\bar{y} = \frac{1}{M}\sum_{m=1}^{M} y_m \tag{11}$$

is the mean prediction across views having Predictions as follows:

$$\mathrm{Var}(y) > \tau_v \tag{12}$$

are discarded. These filters ensure that only high-confidence, low-variance pseudo-labels contribute to training, significantly improving the learning signal from unlabeled documents. To further enhance pseudo-label reliability and the model's robustness across diverse layouts, we introduce cross-view query consistency. This mechanism ensures that object queries remain semantically aligned across different augmented views, reinforcing detection of invariant structural features. Enforcing consistency across views improves generalization to unseen document layouts by aligning predictions under format, scale, and spatial variations.

3.4. Loss Function

As shown in Figure. 1, our loss function combines a supervised loss $\mathcal{L}_s$ on labeled data and an unsupervised loss $\mathcal{L}_u$ on pseudo-labeled data to enhance document layout analysis:

$$\mathcal{L} = \mathcal{L}_s + \lambda\mathcal{L}_u \tag{13}$$

Here, $\mathcal{L}_s$ is calculated using the DETR classification and regression objectives in the box of bounds on the labeled data. The unsupervised loss $\mathcal{L}_u$ is computed on unlabeled data using pseudo-labels generated by the teacher model under weak augmentation. The weighting coefficient λ balances the contributions unsupervised loss.

Method	Epochs	Queries	10% Label			30% Label			50% Label		
			mAP	AP^{50}	AP^{75}	mAP	AP^{50}	AP^{75}	mAP	AP^{50}	AP^{75}
SAM-semi [57]	120	30	92.3	93.7	93.8	93.5	94.8	93.7	93.8	94.8	94.8
our	12	30 + Adaptive	93.1	98.8	96.6	95.5	98.9	98.0	95.5	99.4	98.7

Table 5. Comparison of model performance on the PubTables using 10%, 30%, and 50% labeled data. mAP, AP^{50}, and AP^{75} are evaluated for SAM-semi, and our model.

Method	Epochs	Queries	10% Label	30% Label	50% Label
			AR_L	AR_L	AR_L
SAM-semi [57]	120	30	87.8	88.1	88.3
our	12	30 + Adaptive	95.3	97.2	97.5

Table 6. Comparison of AR_L values between the SAM-semi and our approach on PubTables with 10%, 30%, and 50% labeled data.)

Dataset	Labels	mAP	AP^{50}	AP^{75}	AR_L
PubTables	10%	93.1	98.8	96.6	95.3
	30%	95.5	98.9	98.0	97.2
	50%	95.5	99.4	98.7	97.5

Table 7. Evaluation results for the PubTables dataset using three different data splits: 10%, 30%, and 50%.

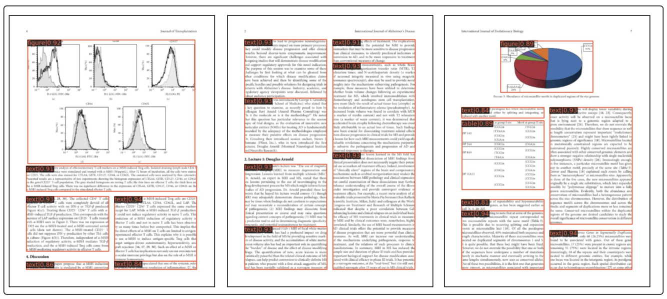

Figure 3. Visualization of detection predictions produced by our approach on the PubLayNet. For a clearer view, please zoom in.

4. Experiments

Dataset: We use three benchmark datasets: PubLayNet [82], PubTables-1M [62], and Table-Bank [28] to evaluate our method. PubLayNet [82] is a large-scale dataset of scientific article pages, comprising more than 360 000 images, annotated with bounding boxes (and polygonal masks) for layout elements like text blocks, figures, and tables, created by automatically aligning PubMed Central PDF and XML sources. PubTables-1M [62] comprises nearly one million tables extracted from scholarly articles, with fine-grained annotations for table detection, structure (rows, columns, cells), and functional analysis, and includes a canonicalization step to address ground-truth inconsistencies. TableBank [28] offers 417 234 high-quality, image-based table annotations gathered via weak supervision from Word and LaTeX documents on the web, enabling both table detection and recognition research.

Implementation and Evaluation Details: Our model is implemented with a ResNet-50 backbone. Training was performed on two NVIDIA RTXA6000 GPUs with a batch size of 16. We employed the AdamW optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 0.0001. Gradient clipping with a maximum norm of 0.1 was applied to stabilize training. The training schedule spanned 12 epochs with a step-based learning rate policy. A pseudo-label confidence threshold of 0.4 was applied to filter uncertain predictions. The teacher model was updated via exponential moving average (EMA) of the student weights with a decay rate of 0.999. We compute mean Average Precision (mAP) across IoU thresholds from 0.50 to 0.95 in increments of 0.05, following COCO-style [32] metrics. Performance on the validation set is also reported using AP^{50}, AP^{75}, and AR_L. Figure 3 illustrates the results of our semi-supervised method on the PubLayNet dataset. These results illustrate the robustness and effectiveness of our approach across a wide range of document layouts.

4.1. Results and Discussion

4.2. PubLayNet

This section evaluates the performance of various models on the PubLayNet dataset using only the table class. For a fair comparison with prior semi-supervised work, annotation files were filtered to include only table instances during training. The analysis focuses on different labeled data splits: 10%, 30%, and 50%. Table 1 presents a comparison of model performance on the table class using 10%, 30%, and 50% labeled data. Our model achieves a mean Average Precision (mAP) of 96.2 at 10% labeled data, significantly outperforming Deformable-semi [53] (88.4) and SAM-semi [57] (89.9) and is competitive with I. Ehsan [16] (97.4). I.Ehsan's approach utilizes a significantly larger number of object queries—specifically, 30 fixed queries combined with 400 modulated ones. In contrast, our method employs a more efficient query strategy with 30 fixed queries and an adaptive number of additional queries ranging from 3 to 9, determined dynamically based on input complexity. Despite the reduced query count, our model achieves competitive results. At 30% labeled data, our model achieves a strong mAP of 96.2, outperforming Deformable-semi (90.3) and SAM-semi (92.8), and showing competitive performance compared to I. Ehsan's model (97.9). At 50% labeled data, with an mAP

Method	Epochs	Queries	10% Label			30% Label			50% Label		
			mAP	AP^{50}	AP^{75}	mAP	AP^{50}	AP^{75}	mAP	AP^{50}	AP^{75}
Deformable-semi [53]	150	30	63.7	93.5	71.6	82.8	96.4	93.4	85.3	96.2	94.4
SAM-semi [57]	120	30	91.2	97.6	96.4	93.7	97.3	96.3	94.8	97.9	97.0
I. Ehsan [16]	150	30 + 400	78.0	87.5	85.5	88.0	94.3	93.1	90.9	97.2	95.9
our	12	30 + Adaptive	91.4	96.6	95.4	92.9	97.0	95.6	93.0	96.6	95.5

Table 8. Performance comparison on the Table-Bank dataset with 10%, 30%, and 50% labeled data, evaluated using mAP, AP^{50}, and AP^{75} for Deformable-semi, SAM-semi, I. Ehsan, and our model.

Method	Epochs	Queries	10% Label	30% Label	50% Label
			AR_L	AR_L	AR_L
Deformable-semi [53]	150	30	74.3	89.0	91.4
SAM-semi [57]	120	30	95.3	97.7	98.1
I. Ehsan [16]	150	30 + 400	90.9	94.1	95.9
our	12	30 + Adaptive	94.9	96.0	96.3

Table 9. Comparison of Average Recall Loss (AR_L) on the Table-Bank dataset, featuring Deformable-semi, SAM-semi, I. Ehsan, and our model.

Dataset	Labels	mAP	AP^{50}	AP^{75}	AR_L
Table-Bank	10%	91.4	96.6	95.4	94.9
	30%	92.9	97.0	95.6	96.0
	50%	93.0	96.6	95.5	96.3

Table 10. Comparison of Average Recall Loss (AR_L) on the Table-Bank dataset, featuring I. Ehsan, and our model.

of 95.9, our model continues to surpass Deformable-semi (90.9) and SAM-semi (93.2), while being slightly below I. Ehsan's higher mAP of 98.6.

Table 2 presents the Average Recall (AR_L) results, where our model consistently outperforms Deformable-semi [53], SAM-semi [57], and I.Ehsan [16] across all labeling percentages. At 10% labeled data, our model achieves an AR_L of 98.2, compared to 91.0, 96.6, and 97.6 for Deformable-semi, SAM-semi, and I.Ehsan, respectively. At 30%, our model achieves 98.3 AR_L, surpassing Deformable-semi's 93.2 and SAM-semi's 96.0, and slightly outperforming I. Ehsan's 98.1. At 50%, our model maintains strong performance with an AR_L of 98.0, higher than Deformable-semi's 96.6 and SAM-semi's 97.3. Although I. Ehsan [16]. slightly surpass our performance in AR_L at 50% labeled data (99.2 vs. 98.0), our model maintains competitive recall while being more efficient.

In addition to comparisons with recent semi-supervised approaches, we include a direct evaluation against fully supervised baselines in Table 4. Notably, even with only 10% labeled data, our semi-supervised model achieves an mAP of 96.2, significantly outperforming the fully supervised methods of Ren et al. [45] (83.4) and Zhu et al. [85] (83.9). This trend continues at 30% and 50% labeled data, where our model achieves mAP scores of 96.2 and 95.9, respectively, outperforming the supervised baselines, which remain below 89. These results demonstrate a clear advantage of our approach over traditional supervised detectors trained with the same amount of labeled data.

As shown in Figure 2, our model demonstrates consistent performance improvement compared to the other methods, highlighting a steady increase in precision across the labeled datasets. Overall, in Table 3, our model consistently achieves high mAP, AP^{50}, AP^{75}, and AR_L scores across 10%, 30%, and 50% labeled data, demonstrating its effectiveness with limited annotations on PubLayNet.

4.3. PubTables-1M

This section evaluates the performance of different models on the PubTables-1M dataset, focusing on labeled data splits of 10%, 30%, and 50%. Table 5 compares our model with SAM-semi [57]. Our model consistently outperforms SAM-semi across all splits, achieving a mAP of 93.1 at 10% labeled data, compared to 92.3 for SAM-semi. At 30%, our model further improves to 95.5 mAP, exceeding SAM-semi's 93.5, demonstrating stronger performance with moderate supervision. At 50%, our method maintains the same mAP of 95.5, still surpassing SAM-semi's 93.8, highlighting its robustness and efficiency even with increased labeled data.

Table 6 shows AR_L results on PubTables-1M, where our model consistently outperforms SAM-semi [57] across all label splits. At 10% labeled data, our model achieves 95.3 AR_L compared to 87.8 for SAM-semi. This advantage persists at 30% and 50%, where our method reaches 97.2 and 97.5, while SAM-semi scores 88.1 and 88.3, respectively.

Table 7 shows evaluation results on the PubTables dataset with 10%, 30%, and 50% labeled data. The model achieves mAP scores of 93.1, 95.5, and 95.5, respectively, with AP^{50} and AP^{75} consistently above 96.6. Average Recall Loss (ARL) improves from 95.3 at 10% to 97.5 at 50%, indicating better performance with more labeled data.

4.4. Table-Bank

This section delves into the performance of various models evaluated on the Table-Bank dataset, focusing on labeled data splits of 10%, 30%, and 50%. Table 8 compares the performance of the Deformable-semi [53], SAM-semi [57], and I. Ehsan [16] methods. At 10% labeled data, our model

o2o	o2m	epochs	Queries	Query Count	mAP	AP^{50}	AP^{75}	AR_L
✓	✗	150	30	Fixed Queries	63.7	93.5	71.6	74.3
✗	✓	120	30	Fixed Queries	91.2	97.6	96.4	95.3
✓	✓	150	30+400	Fixed Queries	78.0	87.5	85.5	90.9
✓	✓	**12**	**30+(3-9)**	**Adaptive Queries**	**91.4**	**96.6**	**95.4**	**94.9**

Table 11. Impact of matching strategies (one-to-one, one-to-many, and hybrid) on detection performance, demonstrating the hybrid approach's ability to achieve the best balance between precision and recall.

outperforms all baselines by achieving a mAP of 92.1, surpassing SAM-semi's 91.2, Deformable-semi's 90.6, and I. Ehsan's 91.3. At 30% labeled data, our model maintains its lead with a mAP of 93.0, closely following SAM-semi's 93.7, and outperforming Deformable-semi's 82.8 and I. Ehsan's 88.0. At 50% labeled data, our model achieves a mAP of 93.0, remaining competitive while SAM-semi leads slightly at 94.8, and outperforming Deformable-semi's 85.3 and I. Ehsan's 90.9.

Table 9 presents the Average Recall Loss (AR_L) results across the 10%, 30%, and 50% labeled data splits. At 10% labeled data, our model achieves an AR_L of 94.9, outperforming Deformable-semi's 74.3 and I. Ehsan's 90.9, while slightly trailing SAM-semi's 95.3. At 30% labeled data, our model maintains strong performance with an AR_L of 96.0, surpassing Deformable-semi (89.0) and I. Ehsan (94.1), though it remains slightly lower than SAM-semi's 97.7. At 50% labeled data, our model records an AR_L of 96.3, again outperforming Deformable-semi (91.4) and I. Ehsan (95.9), while SAM-semi achieves the slightly high score of 98.1. These results demonstrate that our model consistently retains relevant information more effectively than most baselines across varying levels of labeled data.

Table 10 summarizes the overall evaluation results for the Table-Bank data set, showing the mean mean precision (mAP), the average precision at the 0.50 and 0.75 thresholds, and the average recall loss (AR_L) across the three labeled data splits. The results reinforce the efficacy of our model in leveraging labeled data effectively, achieving high precision rates across the board.

5. Ablation Study

This section examines the key design components. All ablation studies are conducted using 10% of the labeled images from the TableBank dataset.

Adaptive Query Strategy This section evaluates the performance of fixed and adaptive query strategies to assess their impact on detection accuracy. As shown in Table 11, the fixed query method (30 + 400) achieves a mean average precision (mAP) of 78.0 and an average recall (AR_L) of 90.9. In comparison, the adaptive query strategy (30 + [3-9]) significantly outperforms the fixed approach, achieving an mAP of 91.4 and an AR_L of 94.9. By dynamically adjusting the number of queries based on the complexity of each document, the adaptive strategy enables more precise detections while minimizing unnecessary computational overhead. These findings highlight the advantages of flexibility in query strategies, resulting in improved detection quality and enhanced model efficiency compared to fixed query methods.

Matching Strategies Table 11 compares one-to-one, one-to-many, and hybrid matching strategies. One-to-one matching with 30 queries achieves a mAP of 63.7 but may miss some objects, as reflected in an AP^{75} of 71.6. One-to-many matching with 30 queries boosts performance to a mAP of 91.2 and AP^{75} of 96.4, improving recall but increasing false positives. A hybrid approach combining one-to-one with 400 queries and one-to-many with 30 queries achieves a mAP of 78.0 and AP^{75} of 85.5, balancing precision and recall. The adaptive strategy, with 30 fixed queries and 3–9 pseudo-label queries, achieves the highest performance with a mAP of 91.4, AP^{50} of 96.6, and AP^{75} of 95.4, demonstrating the best overall detection performance.

6. Conclusion

We present an innovative semi-supervised method that utilizes Focused Attention for Document Layout Analysis. The integration of focused attention allows the model to efficiently prioritize relevant features. We propose a hybrid query mechanism using 30 fixed queries for precise one-to-one matching and 3–9 adaptive queries for broader one-to-many matching, improving detection coverage and accuracy. The Guided Queries mechanism further improves robustness against noise and irrelevant information. This framework employs student and teacher modules, which iteratively refine each other's outputs using the Exponential Moving Average (EMA) function to ensure accurate classification and bounding box predictions. Our experimental results demonstrate that this approach outperforms existing models across multiple benchmark datasets, including PubLayNet, PubTables-1M, and TableBank. Future research could focus on expanding this method to detect other document elements or applying it to diverse document types to further enhance the versatility and generalization of the model.

Acknowledgements

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

References

[1] Sajjad Abbasi, Mohsen Hajabdollahi, Nader Karimi, and Shadrokh Samavi. Modeling teacher-student techniques in deep neural networks for knowledge distillation, 2019. 2, 3

[2] Muhamamd Zeshan Afzal, SK Aziz Ali, Didier Stricker, Peter Eisert, Anna Hilsmann, Daniel Perez-Marcos, Marco Bianchi, Sonia Crottaz-Herbette, Roberto De Ioris, Eleni Mangina, Mirco Sanguineti, Ander Salaberria, Oier Lopez de Lacalle, Aitor García-Pablos, and Montse Cuadros. Next generation xr systems—large language models meet augmented and virtual reality. *IEEE Computer Graphics and Applications*, 45(1): 43–55, 2025. 1, 2

[3] Apostolos Antonacopoulos, David Bridson, Christos Papadopoulos, and Stefan Pletschacher. A realistic dataset for performance evaluation of document layout analysis. In *2009 10th International Conference on Document Analysis and Recognition*, pages 296–300, 2009. 1

[4] Ayan Banerjee, Sanket Biswas, Josep Lladós, and Umapada Pal. Semidocseg: Harnessing semi-supervised learning for document layout analysis, 2023. 1, 3

[5] Galal M. Binmakhashen and Sabri A. Mahmoud. Document layout analysis: A comprehensive survey. *ACM Comput. Surv.*, 52(6), 2019. 1

[6] Galal M. Binmakhashen and Sabri A. Mahmoud. Document layout analysis: A comprehensive survey. *ACM Comput. Surv.*, 52(6), 2019. 1

[7] Jan Brand, Yin-Tat Lee, Danupon Nanongkai, Richard Peng, Thatchaphol Saranurak, Aaron Sidford, Zhao Song, and Di Wang. Bipartite matching in nearly-linear time on moderately dense graphs. pages 919–930, 2020. 1

[8] Gekko Budiutama, Shunsuke Daimon, Hirofumi Nishi, Ryui Kaneko, Tomi Ohtsuki, and Yu-ichiro Matsushita. Channel attention for quantum convolutional neural networks. *Physical Review A*, 110(1), 2024. 3, 4

[9] Kai Chen, Mathias Seuret, Jean Hennebert, and Rolf Ingold. Convolutional neural networks for page segmentation of historical document images. pages 965–970, 2017. 2

[10] Yufan Chen, Jiaming Zhang, Kunyu Peng, Junwei Zheng, Ruiping Liu, Philip Torr, and Rainer Stiefelhagen. Rodla: Benchmarking the robustness of document layout analysis models, 2024. 2, 3

[11] Andrea Corbelli, Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. Historical document digitization through layout analysis and deep content classification. pages 4077–4082, 2016. 1

[12] Andrea Corbelli, Lorenzo Baraldi, Fabrizio Balducci, Costantino Grana, and Rita Cucchiara. Layout analysis and content classification in digitized books. In *Digital Libraries and Multimedia Archives*, pages 153–165, Cham, 2017. Springer International Publishing. 1

[13] Bertrand Coüasnon and Aurélie Lemaitre. Recognition of tables and forms. In *Handbook of Document Image Processing and Recognition*, 2014. 2

[14] Cheng Da, Chuwei Luo, Qi Zheng, and Cong Yao. Vision grid transformer for document layout analysis, 2023. 3

[15] Ana Costa e Silva, Alípio Mário Jorge, and Luís Torgo. Design of an end-to-end method to extract information from tables. *International Journal of Document Analysis and Recognition (IJDAR)*, 8:144–171, 2006. 1, 2

[16] Iqraa Ehsan, Tahira Shehzadi, Didier Stricker, and Muhammad Zeshan Afzal. End-to-end semi-supervised approach with modulated object queries for table detection in documents. page 363–378, Berlin, Heidelberg, 2024. Springer-Verlag. 5, 6, 7

[17] David W. Embley, Matthew F. Hurst, Daniel P. Lopresti, and George Nagy. Table-processing paradigms: a research survey. *International Journal of Document Analysis and Recognition (IJDAR)*, 8:66–86, 2006. 1, 2

[18] Rongyao Fang, Peng Gao, Aojun Zhou, Yingjie Cai, Si Liu, Jifeng Dai, and Hongsheng Li. Feataug-detr: Enriching one-to-many matching for detrs with feature augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–14, 2024. 1

[19] Khurram Azeem Hashmi, Alain Pagani, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. Cascade network with deformable composite backbone for formula detection in scanned document images. *Applied Sciences*, 11(16), 2021. 1, 2

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018. 2

[21] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking, 2022. 1, 2, 3

[22] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 3

[23] Mahmoud Kasem, Abdelrahman Abdallah, Alexander Berendeyev, Ebrahem Elkady, Mohamed Mahmoud, Mahmoud Abdalla, Mohamed Hamada, Sebastiano Vascon, Daniyar Nurseitov, and Islam Taj-Eddin. Deep learning for table detection and structure recognition: A survey. *ACM Comput. Surv.*, 2024. Just Accepted. 2

[24] Muhammad Saif Ullah Khan, Tahira Shehzadi, Rabeya Noor, Didier Stricker, and Muhammad Zeshan Afzal. Enhanced bank check security: Introducing a novel dataset and transformer-based approach for detection and verification, 2024. 1

[25] Andreas Kölsch, Muhammad Zeshan Afzal, Markus Ebbecke, and Marcus Liwicki. Real-time document image classification using deep cnn and extreme learning machines. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pages 1318–1323, 2017. 2

[26] Eunji Lee, Jaewoo Park, Hyung Il Koo, and Nam Ik Cho. Deep-learning and graph-based approach to table structure recognition. *Multimedia Tools Appl.*, 81(4):5827–5848, 2022. 2

[27] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. Dit: Self-supervised pre-training for document image transformer, 2022. 1, 2

[28] Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. Tablebank: A benchmark dataset for table detection and recognition, 2020. 6

[29] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. Docbank: A benchmark dataset for document layout analysis, 2020. 1

[30] Shuai Li, Li Minghan, Ruihuang Li, Chenhang He, and Lei Zhang. One-to-few label assignment for end-to-end dense detection. pages 7350–7359, 2023. 1

[31] Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. Structext: Structured text understanding with multi-modal transformers. pages 1912–1920, 2021. 2

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 6

[33] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018. 2

[34] Mohammad Minouei, Khurram Azeem Hashmi, Mohammad Reza Soheili, Muhammad Zeshan Afzal, and Didier Stricker. Continual learning for table detection in document images. *Applied Sciences*, 12(18), 2022. 1, 2

[35] Mark Niklas Müller, Gleb Makarchuk, Gagandeep Singh, Markus Püschel, and Martin Vechev. Prima: general and precise neural network certification via scalable convex hull approximations. *Proceedings of the ACM on Programming Languages*, 6(POPL):1–33, 2022. 3

[36] Danish Nazir, Khurram Azeem Hashmi, Alain Pagani, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. Hybridtabnet: Towards better table detection in scanned document images. *Applied Sciences*, 11(18), 2021. 2

[37] Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S. Nassar, and Peter Staar. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 3743–3751, New York, NY, USA, 2022. Association for Computing Machinery. 1

[38] Prashanth Pillai and Purnaprajna Mangsuli. Document layout analysis using detection transformers. 2021. 1, 2

[39] Marius-Constantin Popescu, Valentina Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8, 2009. 4

[40] Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. Going full-tilt boogie on document understanding with text-image-layout transformer, 2021. 3

[41] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning, 2017. 2, 3

[42] Sriram Ravichandra, Siva Sathya Sathya, and Lourdumarie Sophie. S. *Deep Learning Based Document Layout Analysis on Historical Documents*, pages 271–281. 2022. 1, 2

[43] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. In *Advances in Neural Information Processing Systems*, pages 29935–29948. Curran Associates, Inc., 2021. 3

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. 1, 2

[45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. 5, 7

[46] Wajahat Saeed, Muhammad Sohail Saleh, Muhammad Naqash Gull, Hassan Raza, Rafaqat Saeed, and Tahira Shehzadi. Geometric features and traffic dynamic analysis on 4-leg intersections. *International Review of Applied Sciences and Engineering*, 2023. 2

[47] Ranajit Saha, Ajoy Mondal, and C. V. Jawahar. Graphical object detection in document images, 2020. 2

[48] Shalini Sarode, Muhammad Saif Ullah Khan, Tahira Shehzadi, Didier Stricker, and Muhammad Zeshan Afzal. Classroom-inspired multi-mentor distillation with adaptive learning strategies, 2025. 2

[49] Sebastian Schreiber, Stefan Agne, Ivo Wolf, Andreas Dengel, and Sheraz Ahmed. 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)*, pages 1162–1167, 2017. 2

[50] Sebastian Schreiber, Stefan Agne, Ivo Wolf, Andreas Dengel, and Sheraz Ahmed. 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)*, pages 1162–1167, 2017. 2

[51] Tahira Shehzadi, Abdul Majid, Madiha Hameed, Aqeel Farooq, and Aqsa Yousaf. 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)*, pages 1–5, 2020. 2

[52] Tahira Shehzadi, Khurram Azeem Hashmi, Alain Pagani, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. Mask-aware semi-supervised object detection in floor plans. *Applied Sciences*, 12(19), 2022. 3

[53] Tahira Shehzadi, Khurram Azeem Hashmi, Didier Stricker, Marcus Liwicki, and Muhammad Zeshan Afzal. Towards end-to-end semi-supervised table detection with deformable transformer. In *Document Analysis and Recognition - ICDAR 2023*, pages 51–76, Cham, 2023. Springer Nature Switzerland. 5, 6, 7

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

[55] Tahira Shehzadi, Khurram Azeem Hashmi, Didier Stricker, Marcus Liwicki, and Muhammad Zeshan Afzal. Bridging the performance gap between detr and r-cnn for graphical object detection in document images. *arXiv preprint arXiv:2306.13526*, 2023. 3

[56] Tahira Shehzadi, Ifza, Didier Stricker, and Muhammad Zeshan Afzal. Semi-supervised object detection: A survey on progress from cnn to transformer, 2024. 3

[57] Tahira Shehzadi, Shalini Sarode, Didier Stricker, and Muhammad Zeshan Afzal. Towards end-to-end semi-supervised table

detection with semantic aligned matching transformer. In *Document Analysis and Recognition - ICDAR 2024: 18th International Conference, Athens, Greece, August 30–September 4, 2024, Proceedings, Part V*, page 295–318, Berlin, Heidelberg, 2024. Springer-Verlag. 5, 6, 7

[58] Tahira Shehzadi, Didier Stricker, and Muhammad Zeshan Afzal. A hybrid approach for document layout analysis in document images, 2024. 2

[59] Tahira Shehzadi, Didier Stricker, and Muhammad Zeshan Afzal. Semitabdetr: End-to-end semi-supervised table detection with transformer-based enhanced query approach, 2024. 3

[60] Tahira Shehzadi, Ifza Ifza, Didier Stricker, and Muhammad Zeshan Afzal. Fd-ssd: Semi-supervised detection of bone fenestration and dehiscence in intraoral images. In *Medical Image Understanding and Analysis*, pages 3–14, Cham, 2026. Springer Nature Switzerland. 3

[61] Gurpreet Singh and Manoj Sachan. Multi-layer perceptron (mlp) neural network technique for offline handwritten gurmukhi character recognition. In *2014 IEEE International Conference on Computational Intelligence and Computing Research*, pages 1–5, 2014. 4

[62] Brandon Smock, Rohith Pesala, and Robin Abraham. Pubtables-1m: Towards comprehensive table extraction from unstructured documents, 2021. 6

[63] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection, 2020. 1, 3

[64] Liang Tang, Tao Li, Yexi Jiang, and Zhiyuan Chen. Dynamic query forms for database queries. *IEEE Transactions on Knowledge and Data Engineering*, 26(9):2166–2178, 2014. 2

[65] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, 2018. 4

[66] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision, 2021. 4

[67] Hong Tai Tran, Nam Quan Nguyen, Tuan Anh Tran, Xuan Toan Mai, and Quoc Thang Nguyen. A deep learning-based system for document layout analysis. In *Proceedings of the 2022 6th International Conference on Machine Learning and Soft Computing*, page 20–25, New York, NY, USA, 2022. Association for Computing Machinery. 1

[68] Daniel Alexander van Strien, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. Assessing the impact of ocr quality on downstream nlp tasks. In *International Conference on Agents and Artificial Intelligence*, 2020. 1

[69] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 2, 3

[70] Jilin Wang, Michael Krumdick, Baojia Tong, Hamima Halim, Maxim Sokolov, Vadym Barda, Delphine Vendryes, and Chris Tanner. A graphical approach to document layout analysis, 2023. 3

[71] Jiawei Wang, Kai Hu, and Qiang Huo. Dlaformer: An end-to-end transformer for document layout analysis. In *Document Analysis and Recognition - ICDAR 2024*, pages 40–57, Cham, 2024. Springer Nature Switzerland. 1

[72] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks, 2020. 3, 4

[73] Xiao Wang and Guo-Jun Qi. Contrastive learning with stronger augmentations, 2022. 3

[74] Yu-An Wang and Yun-Nung Chen. What do position embeddings learn? an empirical study of pre-trained language model positional encoding, 2020. 2

[75] Huichen Yang and William H. Hsu. Vision-based layout detection from scientific literature using recurrent convolutional neural networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6455–6462, 2021. 2

[76] Aqsa Yousaf, Tahira Shehzadi, Aqeel Farooq, and Komal Ilyas. Protein active site prediction for early drug discovery and designing. *International Review of Applied Sciences and Engineering*, 13(1):98 – 105, 2021. 1

[77] Richard Zanibbi, Dorothea Blostein, and James Cordy. A survey of table recognition. *IJDAR*, 7:1–16, 2004. 1

[78] 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, 2022. 2, 3, 4

[79] 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, 2023. 1, 3

[80] Zhenyao Zhao, Min Jiang, Shihui Guo, Zhenzhong Wang, Fei Chao, and Kay Chen Tan. Improving deep learning based optical character recognition via neural architecture search. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, page 1–7. IEEE Press, 2020. 1

[81] Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Weak augmentation guided relational self-supervised learning, 2024. 3

[82] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis, 2019. 6

[83] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Anima Anandkumar, Jiashi Feng, and Jose M. Alvarez. Understanding the robustness in vision transformers, 2022. 2, 3

[84] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications, 2021. 3

[85] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection, 2021. 5, 7

[86] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D. Cubuk, and Quoc V. Le. Rethinking pre-training and self-training, 2020. 1, 3