A Hybrid Approach for Document Layout Analysis in Document images

Tahira Shehzadi*[1,2,3][0000−0002−7052−979X], Didier Stricker[1,2,3], and Muhammad Zeshan Afzal[1,2,3][0000−0002−0536−6867]

[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), 67663 Kaiserslautern, Germany
{tahira.shehzadi@dfki.de}

Abstract. Document layout analysis involves understanding the arrangement of elements within a document. This paper navigates the complexities of understanding various elements within document images, such as text, images, tables, and headings. The approach employs an advanced Transformer-based object detection network as an innovative graphical page object detector for identifying tables, figures, and displayed elements. We introduce a query encoding mechanism to provide high-quality object queries for contrastive learning, enhancing efficiency in the decoder phase. We also present a hybrid matching scheme that integrates the decoder's original one-to-one matching strategy with the one-to-many matching strategy during the training phase. This approach aims to improve the model's accuracy and versatility in detecting various graphical elements on a page. Our experiments on PubLayNet, DocLayNet, and PubTables benchmarks show that our approach outperforms current state-of-the-art methods. It achieves an average precision of **97.3%** on PubLayNet, **81.6%** on DocLayNet, and **98.6%** on PubTables, demonstrating its superior performance in layout analysis. These advancements not only enhance the conversion of document images into editable and accessible formats but also streamline information retrieval and data extraction processes.

Keywords: Detection Transformer · Document Layout Analysis · Graphical object detection

1 Introduction

Systems for Document Intelligence (DI) is essential in enhancing the efficiency of automating large-scale document processing tasks, primarily focusing on extracting and understanding content within these documents. These systems are pivotal in key business intelligence operations such as document retrieval, text recognition, and content categorization, which rely heavily on extracting information and transforming documents into a structured, machine-readable format. This process seamlessly integrates the information extracted into further document

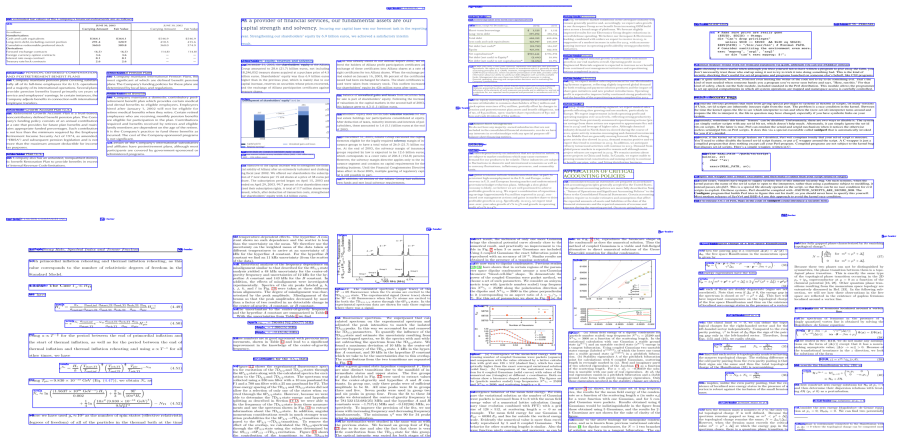

Fig. 1: Diverse layouts and element types in the DocLayNet Dataset, including elements such as captions, footnotes, formulas, and more. It underscores the challenges in document layout analysis, like interpreting dense text and categorizing diverse elements.

processing workflows. As a result, significant improvements have been achieved across various industries, including banking, finance, and healthcare [1,2]. Document Layout Analysis (DLA) has become a key component in Document Intelligence due to its deriving structured formats from unstructured documents. This structuring is vital for accurately identifying and extracting essential document data. DLA encompasses two primary aspects: physical layout analysis, which identifies and spatially categorizes physical page elements like text, images, and tables, and logical layout analysis, which assigns semantic roles to these elements, such as titles, paragraphs, and headers, while understanding their hierarchical and reading order relationships. This analysis is essential for converting scanned documents into editable and searchable formats. However, it faces challenges due to the diversity of document layouts, the varying sizes and shapes of elements, and the complexity of accurately interpreting these elements across different documents.

Previously, remarkable progress has been made in document layout analysis through deep learning techniques, including advanced technologies like Faster RCNN [3] and Mask RCNN [4], as well as other specialized frameworks [5,6]. These methods, effective in specific scenarios such as table detection and the layout analysis of academic papers [7,8,9,10,11], sometimes face limitations in wider applications across various tasks [12]. The advancement of Transformer-based networks [13,14,15,16,17,18,19,20] marks a significant advancement over traditional convolutional neural networks (CNNs), primarily due to their global attention mechanisms and Non-Maximum Suppression (NMS) free design. However, these models still show constraints in precisely detecting textual regions, es-

pecially in identifying small-scale text areas such as headers, footers, and section titles. For example, DINO [19], a leading Transformer-based detection model [19], experiences a notable drop in detection accuracy for these small text regions on the DocLayNet dataset [21]. Fig. 1 shows complex layouts from DocLayNet, with details like captions and footnotes. To improve DLA, we need better algorithms for handling different documents, from academic papers to magazines.

In this paper, we propose an approach to address the challenges of document layout analysis, focusing on accurately identifying graphical elements within pages, such as tables, figures, and formulas. We employ an advanced Transformer-based object detection network [19], for its exceptional capability in detecting various graphical page objects. Enhancing this capability, we introduce a Query Encoding Strategy to provide high-quality object queries by taking high-level query features from the backbone. These query features provide better predictions for small graphical objects like page headers, footers, and titles, combined with the decoder's original queries to improve overall performance. This mechanism is pivotal for contrastive learning, significantly improving the efficiency of the model's decoder phase and enabling more effective processing of complex document layouts. Furthermore, our approach introduces a novel hybrid matching scheme that merges the decoder's original one-to-one matching strategy with an auxiliary one-to-many matching strategy. This integration, implemented during the training phase, is key to boosting the model's accuracy and adaptability in recognizing diverse classes of graphical elements. By combining the transformer's object detection capabilities with our unique encoding query and selection strategies, our method sets a new benchmark in document layout analysis, significantly advancing the field's ability to accurately detect and interpret graphical elements within various documents.
We summarize the main contributions of this paper as follows:

- We introduce a Transformer-based framework for document layout analysis, incorporating a ResNet-50 backbone. This framework is augmented with an enhanced query encoding mechanism and innovative query-selection strategies. By integrating these strategies, our approach sets a new standard in document layout analysis. This significant advancement contributes to accurately detecting graphical elements in various document types.
- We present a unique query selection scheme that blends the decoder's original one-to-one matching strategy with a one-to-many matching strategy. This integration, crucial during the training phase, significantly enhances the model's accuracy and adaptability in detecting and categorizing various graphical elements across different documents.
- We introduce an enhanced query encoding mechanism to improve the efficiency of the model's decoder phase and enable more effective processing of complex document layouts.
- To validate the effectiveness of our approach, we conduct comprehensive evaluations on three distinct datasets: PubLayNet, DocLayNet, and Pubtables. These evaluations demonstrate the robustness and applicability of our proposed method across various documents and layout challenges.

2 Related Work

Layout analysis is crucial to extract data from digital documents effectively. It involves understanding the spatial arrangement and relationships between various elements like tables, text, figures, and titles. Before deep learning approaches [3,4,22], heuristic rule-based algorithms [23,24] were emplyed in layout analysis. However, with technological advancements, convolutional neural networks (CNNs) became the primary method, providing significant improvements. More recently, transformer-based architectures [13,25,26,27,15,16,28,18,19] have emerged as the leading approach, show remarkable effectiveness in this domain. This section aims to offer an in-depth review of these cutting-edge techniques, exploring a variety of approaches in Document Layout Analysis (DLA).

Heuristic Rule-Based DLA. The document layout analysis using heuristic techniques is generally categorized into top-down, bottom-up, and hybrid approaches. Bottom-up methods [23,24] involve elementary processes such as clustering and combining pixels to form uniform regions for akin objects while segregating dissimilar ones. Conversely, top-down approaches [29,30] iteratively divide the document image into various regions until distinct areas encompassing similar objects are formed. While bottom-up strategies are capable of handling intricate layouts, they require significant computational resources. On the other hand, top-down methods are more efficient in terms of implementation speed but lack versatility, showing optimal performance only with certain document types. Hybrid methods [31,32] combine the strengths of both bottom-up and top-down techniques, achieving both rapid and effective outcomes. Before the advent of deep learning, these heuristic strategies were the leading methods for detecting tables in documents.

Deep Learning-based DLA. With the rise of deep learning approaches [7,8,10], Convolutional Neural Networks (CNNs) have performed better than traditional rule-based algorithms in document analysis [33,34,35,36,37,38,39]. This development represents a significant improvement in the precision and efficiency of processing and understanding complex document layouts. Introducing Faster-RCNN [3] marks a significant advancement in document object detection, facilitating effective page segmentation [40]. Subsequently, Mask-RCNN [4] set a new benchmark in layout segmentation, particularly for newspapers. RetinaNet [22] further contributes to this evolution by focusing on keyword detection in document images, although its complexity limits its application to text region detection. For table detection and structural recognition, DeepDeSRT [7] introduces an innovative image transformation approach that discerns table features for input into a fully convolutional network employing skip pooling. The ICDAR2017 POD (Page Object Detection) benchmark, introduced by Saha et al. [41], utilizes a transfer learning-based Faster-RCNN architecture to detect elements like mathematical equations, tables, and figures. To address cross-domain challenges in Document Object Detection, a new benchmark [42] is established, focusing on domain adaptation strategies. More recently, a vision-based layout detec-

tion benchmark [43] employs a recurrent convolutional neural network with a VoVNet-v2 backbone, generating synthetic PDF documents from the ICDAR-2013 and GROTOAP datasets to set new standards in scientific document analysis.

Transformer-Based DLA. Document layout analysis is rapidly advancing with Transformer architectures, known for their positional embedding and attention mechanisms. These methods are known for their unique features like positional embedding and attention mechanism [44]. DiT [45] has set a new standard in classifying document images, layout analysis, and table detection, employing self-supervised training on extensive collections of unlabeled document images. However, its application is limited to smaller datasets like PRIMA. Li et al. [46] develop a method that combines different types of data to understand structured text in documents. However, this method struggles with text that has similar meanings. Furthermore, the TILT [47] model simultaneously processes textual, visual, and layout data through an encoder-decoder Transformer setup. Another implementation of a transformer encoder-decoder in [48] establishes a benchmark for the PubLayNet dataset [49], integrating text data extracted via OCR. The LayoutLMv3 [50] model improves visual document understanding by jointly learning from text, layout, and visual elements. It performs better with large datasets but has limitations with smaller ones. Other recent models [51,52,53,54] also adopt joint pre-training strategies for various tasks, including document visual question answering. The transformer-based architectures have emerged as the leading approaches, show remarkable effectiveness in object detection domain [13,14,25,26,17,27,15,16,28,55,18,19]. However, when employing DINO [19] or other Transformer-based networks in document layout analysis, there's a noted limitation in their performance with small graphical objects like page titles, headers, and footers. To improve this, We enhance the hybrid query mechanism and matching scheme. This strategy elevates our document layout analysis, allowing for more precise and flexible detection and interpretation of various document graphical elements.

3 Methodology

Our approach consists of four integral parts: First, a CNN-based backbone network for extracting multi-scale features from document images. Second, a transformer-based model is employed to detect graphical elements like titles, figures, tables, and text on the pages. Third, we introduce an improved query encoding mechanism, optimizing the model's decoder phase to process complex document layouts more effectively. Fourthly, we implement a unique query selection scheme, blending the decoder's one-to-one matching with a new one-to-many strategy, enhancing accuracy in identifying various graphical elements during training. These modules are collectively trained in an end-to-end manner. The complete overview of our approach is shown in Fig. 2 and explained in detail in the subsequent subsections.

3.1 Backbone Multi-scale Features Network

For processing an input image I of size H×W×3, we use a ResNet-50 backbone network to generate a series of feature maps at reduced resolutions: 1/4, 1/8, 1/16, 1/32, and 1/64 of the original size. Each map is refined using a 1×1 convolution layer, which is crucial for reducing the channel count. This step is essential to control the number of trainable parameters, making the process manageable, especially with limited computational resources. After this reduction, each feature map has 256 channels, which are then input into the transformer network to detect graphical objects on the page.

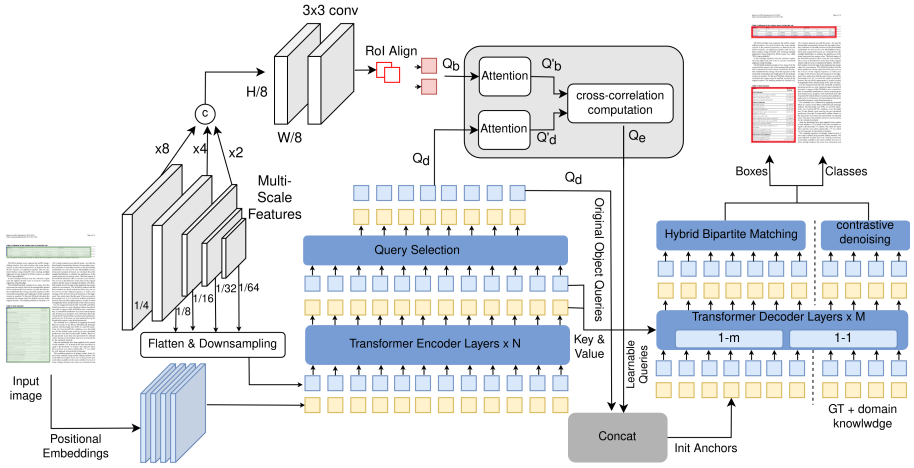

Fig. 2: Overview of our approach for Document Layout Analysis. The input image is processed through a CNN backbone to extract features, which are then passed to a Transformer encoder-decoder network. The encoder processes the features globally, while the decoder uses object queries to interact with the encoded features and predict bounding boxes and classes for each object in the image. Our approach incorporates an enhanced query encoding mechanism to improve decoder efficiency and a query selection scheme that combines one-to-one and one-to-many matching strategies, improving accuracy and adaptability in identifying various graphical elements across documents.

3.2 Document Layout Analysis with the transformer Framework

Recent progress in Transformer-based object detection [20,13,14,56,16,28,18,19] has revolutionized document analysis. These advanced methods outperform previous models like Faster-RCNN [3] and Mask-RCNN [4], mainly because they don't require manual techniques such as anchor generation process and NMS. Our approach employs the DINO [19] model, a state-of-the-art network, to detect

graphical elements in document images. Our approach includes a transformer network with a unique structure. It has an encoder that processes variously scaled feature maps from a CNN backbone and a decoder that generates the final results. The encoder's task is to generate detailed proposals for graphical elements in the pages, guiding the decoder's positional embeddings for the queries. This decoder uses a deformable attention mechanism for better efficiency in self and cross-attention processes. It also applies contrastive denoising for the object queries, helping the model learn faster. It is highly adaptable, especially during shifts in document types, and it focuses on lower-dimension image features, which often need more data to be included in traditional transformer training. The effectiveness of our transformer-based approach is validated through its impressive performance in detecting graphical page objects on well-known benchmarks like PubLayNet, DocLayNet, and PubTables.

3.3 Query Encoding Strategy

In the query encoding strategy, we enhance the query mechanism to improve the detection of small graphical objects in document images by combining backbone query features with decoder original queries. This approach creates high-quality object queries, increasing accuracy in identifying the small elements within an image. Here's a detailed explanation:

High-level Query Features from Backbone: In our approach, we initially extract high-level features from the early layers of a CNN backbone, such as the ResNet-50. These initial layers are adept at capturing intricate details and textures, including edges, corners, and specific patterns. This level of granularity is crucial for identifying smaller objects within an image. For each processed image, we adjust the dimensions of feature maps $C4$ and $C5$ to align with $C3$ and then concatenate them. The combined feature map undergoes processing through two 3×3 convolutional layers, resulting in a feature map C_{h} comprising 64 channels. Then, we employ the RoIAlign algorithm [57] to extract features based on its bounding box $b_j = (x_{j1}, y_{j1}, x_{j2}, y_{j2})$ with many MLPs , where (x_{j1}, y_{j1}) and (x_{j2}, y_{j2}) are the coordinates of the upper-left and lower-right corners, respectively. The high-level query features are then defined as:

$$Q_h = MLP(\mathrm{RoIAlign}(F_{\mathrm{h}}, b_j)) \tag{1}$$

Here, Q_h refers to the query features and F_h denotes features from backbone. Next, we apply a self-attention mechanism, as described in [44], to the high-level query features Q_h and decoder orignal query features Q_d. This mechanism enables the model to prioritize and weight the importance of different aspects of the high-level features, thus enhancing the overall feature representation. Following the self-attention step, we determine the cross-correlation (similarity) between the outputs Q'_h and Q'_d, using cosine similarity [58]. The process is formalized as follows:

$$Q_e = similarity(Q'_d, Q'_h) \tag{2}$$

Here, Q'_h refers to the query features from the backbone after self-attention, while Q'_d denotes the original decoder queries after self-attention. The final step involves integrating these refined queries Q_e with the original queries from the decoder, enhancing the overall feature extraction and analysis.

Combining Features for Enhanced Detection: In the next step, we integrate two distinct types of features to boost detection capabilities. High-level features, represented by Q_h, allow the model to understand the overall layout and context of the document. On the other hand, we have the original transformer query features, which are adept at capturing specific object information. The enhanced features Q_e are obtained from self-attention on Q_h and decoder original queries Q_d. This concatenation is particularly beneficial for detecting small graphical elements, such as page headers, footers, and titles, which might be missed by the decoder's original query features alone. Combining these features enhances the model's detection sensitivity to these smaller elements. The combined query features, which we denote as Q_t, are formed by concatenating the decoder original query features with the enhanced query features Q_e:

$$Q_t = \text{Concat}(Q_d, Q_e) \tag{3}$$

This combination of features from both the high-level and the decoder queries enriches the feature set provided to the model, leading to a more robust detection mechanism for various objects within complex documents.

Integration with Decoder's Original Queries: By merging previously generated queries with the original decoder queries, our model performs better in identifying elements in document images. This integration enhances the model's ability to detect prominent and subtle features within complex document layouts, making it especially effective for predicting small or easily overlooked objects. The process of query integration and output generation is formulated as:

$$o = Decoder(Q_t, E|A) \tag{4}$$

Here, our model utilizes a set of decoder queries, denoted by Q_t, and corresponding outputs from the Transformer decoder, represented as o. The refined image features, processed by the Transformer encoder, are symbolized by F, while A represents the attention mask, specifically designed for the denoising task [59]. In this way, this query mechanism combines the strengths of both abstract and detailed image features, facilitating thorough and precise detection of diverse elements within intricate document structures.

3.4 Query Selection Strategy

Our research introduces an innovative hybrid matching scheme for analyzing complex documents. This approach uniquely combines two query strategies, one-to-one and one-to-many matching, to enhance the detection and understanding

of various elements in detailed documents. Initially, we observe that the one-to-many strategy led to duplicate predictions, as shown in Table 5. To optimize this, we utilized one-to-many matching during the first half of our training iterations, then shifted to one-to-one matching for the remainder. This transition markedly improved accuracy and reduced duplications.

As a key feature of our hybrid approach, the one-to-many matching branch is designed to enhance object detection in complex document layouts. This innovative branch enables the association of a single detected object with multiple ground truths, a significant advancement over traditional one-to-one matching methods. We integrate original decoder queries with high-quality object queries generated in the one-to-many strategy. These object queries are generated by merging high-level query features from the backbone as explained in subsection 3.3. It is particularly useful in complex document layouts where traditional one-to-one matching might struggle. By enabling an object to be matched with several ground truths, the model better understands the document's content, especially in overlapping or closely packed elements. The total loss in ono-to-many strategy is as follows:

$$L_{cls}^{1-m} = \sum_{i=1}^{N_{obj}} |\hat{g}_i - p_i| \cdot BCE(p_i, \hat{g}_i) + \sum_{j=1}^{N_{no}} p_j \cdot BCE(p_j, 0) \tag{5}$$

$$L_{reg}^{1-m} = \sum_{i=1}^{N_{obj}} \hat{g}_i \cdot \mathcal{L}_{GIoU}(bx_i, \hat{bx_i}) + \sum_{i=1}^{N_{obj}} \hat{g}_i \cdot \mathcal{L}_{L1}(bx_i, \hat{bx_i}) \tag{6}$$

$$L^{1-m} = L_{cls}^{1-m} + L_{reg}^{1-m} \tag{7}$$

where $\hat{g}_i$ is the ground truth, p_i is the actual prediction. In the one-to-one matching branch, a traditional approach in object detection models, each detected object is directly aligned with a corresponding ground truth. This method is straightforward and effective in scenarios where objects are clearly separated and easily identifiable. It eliminates duplications generated in the one-to-many strategy, ensuring more accurate predictions. The total loss in the one-to-one matching strategy is as follows:

$$L^{1-1} = L_{cls}^{1-1} + L_{reg}^{1-1} \tag{8}$$

This hybrid approach retains the benefits of the traditional method, like eliminating the need for Non-Maximum Suppression (NMS), and does not add any extra computational cost. The combination of these two methods in a single model allows for more accurate and efficient object detection in a wide range of scenarios, significantly improving the performance of document analysis tasks.

4 Experimental Setup

Datasets and Evaluation Criteria. Our study employs three benchmark datasets to evaluate the efficacy of the proposed method: PubLayNet [49] PubTables [60] and DocLayNet [21]. We adopt the mean Average Precision (mAP)

metric in line with COCO-style [61] standards to evaluate our approach. We compute precision across a spectrum of Intersection over Union (IoU) thresholds, from 0.50 to 0.95, increasing in 0.05 steps. This IoU range is essential for evaluating our model's accuracy in category-specific tasks. Our mAP calculation, averaged across these IoU levels, follows the established Microsoft COCO benchmark, facilitating a standardized comparison with other models. We further refine our assessment by calculating Average Precision (AP) at specific IoU thresholds of 0.50 and 0.75, offering a focused analysis of the model's performance at these recognized benchmarks. It clearly explains our model's proficiency in accurately classifying various categories.

Implementation Details. Our network is trained on RTXA600 GPUs, utilizing a ResNet-50 network as the backbone, which is pre-trained on ImageNet. We employ the AdamW algorithm for optimization, with a batch size of 16. The training duration is set to 12 epochs for both PubLayNet and PubTables datasets, and extended to 24 epochs for the DocLayNet dataset. We implement a learning rate reduction strategy, decreasing it by a factor of 10 later in the training process. Our approach includes a multi-scale training technique, where images are resized to various lengths without exceeding a maximum size limit. For the testing phase, we resize images to have a shorter side of 640, optimizing image handling during model evaluation.

Table 1: **Evaluation on the DocLayNet Benchmark.** A comparative analysis of outcomes on the DocLayNet Test Dataset. Here, Mask represents Mask R-CNN and Faster indicates Faster R-CNN. In this comparison, the performances of Mask R-CNN, Faster R-CNN, and YOLOv5 are referenced from [21], and the results for the DINO model are derived from [62]. The best results are highlighted in bold.

Classes	**Mask**	**Faster**	**YOLOv5**	**DINO**	**Zhong et al. [62]**	**Ours**
Caption	71.5	70.1	77.7	85.5	83.2	**85.6**
Footnote	71.8	73.7	**77.2**	69.2	69.7	70.0
Formula	63.4	63.5	**66.2**	63.8	63.4	64.7
List-item	80.8	81.0	86.2	80.9	**88.6**	83.5
Page-footer	59.3	58.9	61.1	54.2	90.0	**91.3**
Page-header	70.0	72.0	67.9	63.7	76.3	**77.8**
Picture	72.7	72.0	77.1	84.1	81.6	**84.7**
Section-header	69.3	68.4	74.6	64.3	**83.2**	82.9
Table	82.9	82.2	86.3	85.7	84.8	**86.1**
Text	**85.8**	85.4	88.1	83.3	84.8	85.4
Title	80.4	79.9	82.7	82.8	84.9	**86.3**
All	73.5	73.4	76.8	74.3	81.0	**81.6**

5 Results and Discussion

5.1 DocLayNet

Table 1 summarizes the performance of our approach compared to other approaches on the DocLayNet dataset, with results measured in mean Average Precision (mAP) for different document elements. Our method outperforms previous networks like Mask R-CNN [57], Faster R-CNN [63], YOLOv5 [64], DINO, and the document analysis approach of Zhong et al. [62], particularly in recognizing 'Caption,' 'Page-footer,' 'Page-header,' and 'Title,' achieving the highest overall mAP at 81.6%. This comprehensive evaluation across various classes highlights the effectiveness of our approach in accurately detecting and classifying elements in a wide array of document layouts.

Fig. 3: Visual analysis of our approach on the DocLayNet dataset. Here, blue color represents ground truth, red denotes prediction by our approach. It illustrates the model's proficiency in identifying small layout elements, specifically highlighting its accuracy in detecting page titles, headers, and footers.

Fig. 3 illustrates the visual results of our document layout analysis approach on the DocLayNet dataset. It displays document page with our model's predictions compared to the actual ground truth (GT). In these visual examples, ground truth annotations are outlined in blue, and our model's predictions are in red. This comparison aims to showcase our model's precision in identifying small layout elements, such as page titles, headers, and footers. Using contrasting colors demonstrates the accuracy of our approach in detecting and classifying the intricate details of document layouts.

5.2 PubLayNet

We also evaluate and compare our approach with previous document analysis approaches on the PubLayNet dataset. The results of these comparisons are detailed in Table 2. The results indicate that our approach significantly outper-

Table 2: **Evaluation on the PubLayNet Benchmark.** A comparative analysis of results on the PubLayNet Validation Set. The results highlight the effectiveness of our approach. The best results are highlighted in bold.

Method	Text	Title	List	Table	Figure	mAP
Faster R-CNN [49]	91.0	82.6	88.3	95.4	93.7	90.2
Mask R-CNN [49]	91.6	84.0	88.6	96.0	94.9	91.0
Naik et al. [11]	94.3	88.7	94.3	97.6	96.1	94.2
Minouei et al. [65]	94.4	90.8	94.0	97.4	96.6	94.6
DiT-L [45]	94.4	89.3	96.0	97.8	97.2	94.9
SRRV [66]	95.8	90.1	95.0	97.6	96.7	95.0
DINO [19]	94.9	91.4	96.0	98.0	97.3	95.5
TRDLU [48]	95.8	92.1	**97.6**	97.6	96.6	96.0
UDoc [53]	93.9	88.5	93.7	97.3	96.4	93.9
LayoutLMv3 [50]	94.5	90.6	95.5	97.9	97.0	95.1
VSR [67]	96.7	93.1	94.7	97.4	96.4	95.7
Zhong et al. [62]	97.4	93.5	96.4	98.2	97.2	96.5
Our	**98.0**	**94.2**	97.3	**98.6**	**98.5**	**97.3**

forms previous methods, demonstrating its superior performance in document analysis.

5.3 PubTables

We also evaluate our approach and compare it with previous table detection approaches on PubTables dataset. The results of these comparisons are detailed in Table 3. The results clearly demonstrate that our approach outperforms previous table detection approaches, highlighting its effectiveness and efficiency in accurately identifying and classifying table elements within complex documents.

Table 3: Comparative Analysis of Results on the PubTables Validation Set. The best results are highlighted in bold.

Method	Detector	mAP	AP^{50}	AP^{75}
Smock et al. [60]	Faster R-CNN	82.5	98.5	92.7
Smock et al. [60]	DETR	96.6	995	98.8
Minouei et al. [8]	Sparse R-CNN+PVT	98.2	-	-
Our	DINO	**98.6**	**99.8**	**99.1**

5.4 Ablation Study

In our ablation study, we explore the impact of object query selection, the effectiveness of matching strategies, and the influence of the quantity of learnable queries in our Transformer-based model. This investigation is designed to observe how these key components individually and collectively affect our model's precision and functionality in analyzing complex document layouts.

Table 4: Detailed Ablation Analysis on the PubLayNet Validation Dataset.

Method	Text	Title	List	Table	Figure	mAP
DINO-Queries (Q_d)	94.9	91.4	96.0	98.0	97.3	95.52
Hybrid-Queries ($Q_d + Q_e$)	98.0	94.2	97.3	98.6	98.5	97.3

Influence of object query selection In the ablation study, we observe the impact of object query selection, which is crucial for detecting small graphical objects like page headers, footers, and titles in document layout analysis. The study examines the enhanced query mechanism that combines high-level backbone features with decoder original query features. By integrating the refined query features with the original queries, we observe a significant improvement, as shown in Table 4, to accurately predict and identify smaller elements within document layouts. This comparison shows how high-quality object queries improve document analysis. It demonstrates that modifying query integration can significantly enhance the model's performance and accuracy.

Influence of matching strategy In our document layout analysis approach, employing one-to-one and one-to-many matching strategies provides a comprehensive approach, as shown in Table 5. In our setup, Q_d represents the standard decoder queries, while Q_e represents the enhanced queries. The data indicates that the best mean Average Precision (mAP) is obtained when these two sets of queries are used together, which leads to better training results. It's particularly effective to start training with Q_d and Q_e and then transition to using just Q_d halfway through the training process. One-to-one matching using Q_d, aligning each prediction with a single ground truth, is efficient for clear, distinct objects,

Table 5: Performance comparison using various query combinations as input to the decoder on PubLayNet dataset. Here, Q_d represents the original decoder queries, while Q_e signifies the enhanced queries. The highest mean Average Precision (mAP) is achieved by combining the original DINO queries with the enhanced queries, indicating improved performance during training with overlap in predictions. A training approach that uses $Q_d + Q_e$ for the initial half of training epochs before switching to Q_d is shown to be effective.

$\mathbf{Q_d}$	$\mathbf{Q_d + Q_e}$	NMS-free	mAP	$\mathbf{AP^{50}}$	$\mathbf{AP^{75}}$
✓	✗	✓	95.5	-	-
✗	✓	✗	98.4	98.8	97.7
✓	✓	✓	97.3	98.5	97.4

ensuring straightforward training. On the other hand, one-to-many matching employing $Q_d + Q_e$ allows a single prediction to correspond to several ground truths, adeptly handling complex layouts with overlapping or closely packed elements. This dual strategy leverages the strengths of both approaches, enhancing the model's ability to accurately detect and classify a wide range of object types in various document layouts.

Influence of Learnable queries Quantity The quantity of learnable queries in Transformer-based models like DINO significantly affects their performance in document layout analysis, as observed in Table 6. More queries enable finer detection of detailed elements and improve overall accuracy, but there's a need for balance. Excessive queries can increase computational demands and risk overfitting, while too few may miss intricate details. Thus, optimizing the number of queries is crucial for efficient processing, balancing computational resources, and ensuring adaptability across various document types and complexities.

Table 6: Performance comparison using different numbers of learnable queries to the decoder input on PubLayNet Dataset. The best-performing results are highlighted in bold, illustrating the optimal number of queries required for best model performance. As indicated, the model generally improves with more queries, up to a point, after which the performance decreases, suggesting an optimal query range for efficient detection across various object sizes.

N	AP	$\mathbf{AP^{50}}$	$\mathbf{AP^{75}}$	$\mathbf{AP_s}$	$\mathbf{AP_m}$	$\mathbf{AP_l}$
100	95.3	96.7	95.8	35.8	65.3	89.2
200	96.5	97.3	96.6	43.5	71.8	96.4
300	**97.3**	**98.5**	**97.4**	**43.8**	**72.7**	**96.7**
400	96.4	98.2	97.0	43.1	60.7	96.1

6 Conclusion

This paper introduces a approach for analyzing document layouts, focusing on accurately identifying elements like text, images, tables, and headings in documents. We introduce a hybrid query mechanism that enhances object queries for contrastive learning, improving the efficiency of the decoder phase in the model. Moreover, during training, our approach features a hybrid matching scheme that combines the decoder's original one-to-one matching with a one-to-many matching branch, aiming to increase the model's accuracy and flexibility in detecting diverse graphical elements on a page. We evaluate our approach on benchmark datasets like PubLayNet, DocLayNet, and PubTables. It demonstrates superior accuracy and precision in layout analysis, outperforming current state-of-the-art methods. These advancements significantly aid in transforming document images into editable and accessible formats, streamlining information retrieval and data extraction processes. The implications of our research are substantial, affecting areas such as digital archiving, automated form processing, and content management systems. This work represents a significant contribution to document analysis and digital information management, setting new benchmarks and paving the way for future advancements.

References

1. L. Cui, Y. Xu, T. Lv, and F. Wei, "Document AI: benchmarks, models and applications," *CoRR*, vol. abs/2111.08609, 2021. [Online]. Available: https://arxiv.org/abs/2111.08609
2. T. Shehzadi, A. Majid, M. Hameed, A. Farooq, and A. 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)*, vol. 5, 2020, pp. 1–5.
3. S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015. [Online]. Available: http://arxiv.org/abs/1506.01497
4. K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *CoRR*, vol. abs/1703.06870, 2017. [Online]. Available: http://arxiv.org/abs/1703.06870
5. Z. Cai and N. Vasconcelos, "Cascade R-CNN: delving into high quality object detection," *CoRR*, vol. abs/1712.00726, 2017. [Online]. Available: http://arxiv.org/abs/1712.00726
6. N. Ma, X. Zhang, H. Zheng, and J. Sun, "Shufflenet V2: practical guidelines for efficient CNN architecture design," *CoRR*, vol. abs/1807.11164, 2018. [Online]. Available: http://arxiv.org/abs/1807.11164
7. S. Schreiber, S. Agne, I. Wolf, A. Dengel, and S. 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)*, vol. 01, 2017, pp. 1162–1167.
8. M. Minouei, K. A. Hashmi, M. R. Soheili, M. Z. Afzal, and D. Stricker, "Continual learning for table detection in document images," *Applied Sciences*, vol. 12, no. 18, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/18/8969

9. S. Sinha, K. A. Hashmi, A. Pagani, M. Liwicki, D. Stricker, and M. Z. Afzal, "Rethinking learnable proposals for graphical object detection in scanned document images," *Applied Sciences*, vol. 12, no. 20, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/20/10578
10. T. Shehzadi, K. A. Hashmi, A. Pagani, M. Liwicki, D. Stricker, and M. Z. Afzal, "Mask-aware semi-supervised object detection in floor plans," *Applied Sciences*, vol. 12, no. 19, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/19/9398
11. S. Naik, K. A. Hashmi, A. Pagani, M. Liwicki, D. Stricker, and M. Z. Afzal, "Investigating attention mechanism for page object detection in document images," *Applied Sciences*, vol. 12, no. 15, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/15/7486
12. L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-dujaili, Y. Duan, O. Al-Shamma, J. I. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, cnn architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:232434552
13. X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable {detr}: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=gZ9hCDWe6ke
14. Z. Dai, B. Cai, Y. Lin, and J. Chen, "Up-detr: Unsupervised pre-training for object detection with transformers," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1601–1610, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:227011943
15. T. Wang, L. Yuan, Y. Chen, J. Feng, and S. Yan, "Pnp-detr: Towards efficient visual analysis with transformers," *CoRR*, vol. abs/2109.07036, 2021. [Online]. Available: https://arxiv.org/abs/2109.07036
16. Y. Fang, B. Liao, X. Wang, J. Fang, J. Qi, R. Wu, J. Niu, and W. Liu, "You only look at one sequence: Rethinking transformer in vision through object detection," *CoRR*, vol. abs/2106.00666, 2021. [Online]. Available: https://arxiv.org/abs/2106.00666
17. T. Shehzadi, K. Azeem Hashmi, D. Stricker, M. Liwicki, and M. Zeshan Afzal, "Towards end-to-end semi-supervised table detection with deformable transformer," in *Document Analysis and Recognition - ICDAR 2023*, G. A. Fink, R. Jain, K. Kise, and R. Zanibbi, Eds. Cham: Springer Nature Switzerland, 2023, pp. 51–76.
18. Z. Chen, J. Zhang, and D. Tao, "Recurrent glimpse-based decoder for detection with transformer," *CoRR*, vol. abs/2112.04632, 2021. [Online]. Available: https://arxiv.org/abs/2112.04632
19. H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," 2022. [Online]. Available: https://arxiv.org/abs/2203.03605
20. T. Shehzadi, K. A. Hashmi, D. Stricker, and M. Z. Afzal, "2d object detection with transformers: A review," *arXiv preprint arXiv:2306.04670*, 2023.
21. B. Pfitzmann, C. Auer, M. Dolfi, A. S. Nassar, and P. 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*, 2022, pp. 3743–3751.
22. T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *CoRR*, vol. abs/1708.02002, 2017. [Online]. Available: http://arxiv.org/abs/1708.02002

23. A. Asi, R. Cohen, K. Kedem, and J. El-Sana, "Simplifying the reading of historical manuscripts," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, pp. 826–830.
24. R. Saabni and J. El-Sana, "Language-independent text lines extraction using seam carving," in *2011 International Conference on Document Analysis and Recognition*, 2011, pp. 563–568.
25. P. Gao, M. Zheng, X. Wang, J. Dai, and H. Li, "Fast convergence of DETR with spatially modulated co-attention," *CoRR*, vol. abs/2101.07448, 2021. [Online]. Available: https://arxiv.org/abs/2101.07448
26. D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, L. Sun, and J. Wang, "Conditional DETR for fast training convergence," *CoRR*, vol. abs/2108.06152, 2021. [Online]. Available: https://arxiv.org/abs/2108.06152
27. F. Liu, H. Wei, W. Zhao, G. Li, J. Peng, and Z. Li, "Wb-detr: Transformer-based detector without backbone," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 2959–2967.
28. W. Wang, Y. Cao, J. Zhang, and D. Tao, "FP-DETR: Detection transformer advanced by fully pre-training," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=yjMQuLLcGWK
29. N. Journet, V. Eglin, J. Ramel, and R. Mullot, "Text/graphic labelling of ancient printed documents," in *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, 2005, pp. 1010–1014 Vol. 2.
30. K. Kise, A. Sato, and M. Iwata, "Segmentation of page images using the area voronoi diagram," *Computer Vision and Image Understanding*, vol. 70, no. 3, pp. 370–382, 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314298906841
31. J. Chen and D. Lopresti, "Table detection in noisy off-line handwritten documents," in *2011 International Conference on Document Analysis and Recognition*, 2011, pp. 399–403.
32. J. Fang, L. Gao, K. Bai, R. Qiu, X. Tao, and Z. Tang, "A table detection method for multipage pdf documents via visual seperators and tabular structures," in *2011 International Conference on Document Analysis and Recognition*, 2011, pp. 779–783.
33. M. Minouei, K. A. Hashmi, M. R. Soheili, M. Z. Afzal, and D. Stricker, "Continual learning for table detection in document images," *Applied Sciences*, vol. 12, no. 18, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/18/8969
34. G. Kallempudi, K. A. Hashmi, A. Pagani, M. Liwicki, D. Stricker, and M. Z. Afzal, "Toward semi-supervised graphical object detection in document images," *Future Internet*, vol. 14, no. 6, 2022. [Online]. Available: https://www.mdpi.com/1999-5903/14/6/176
35. K. A. Hashmi, A. Pagani, M. Liwicki, D. Stricker, and M. Z. Afzal, "Castabdetectors: Cascade network for table detection in document images with recursive feature pyramid and switchable atrous convolution," *Journal of Imaging*, vol. 7, 2021.
36. D. Nazir, K. A. Hashmi, A. Pagani, M. Liwicki, D. Stricker, and M. Z. Afzal, "Hybridtabnet: Towards better table detection in scanned document images," *Applied Sciences*, vol. 11, no. 18, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/18/8396
37. K. A. Hashmi, A. Pagani, M. Liwicki, D. Stricker, and M. Z. Afzal, "Cascade network with deformable composite backbone for formula detection in scanned

document images," *Applied Sciences*, vol. 11, no. 16, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/16/7610

38. K. A. Hashmi, D. Stricker, M. Liwicki, M. N. Afzal, and M. Z. Afzal, "Guided table structure recognition through anchor optimization," *CoRR*, vol. abs/2104.10538, 2021. [Online]. Available: https://arxiv.org/abs/2104.10538
39. A. Kölsch, M. Z. Afzal, M. Ebbecke, and M. 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)*, vol. 01, 2017, pp. 1318–1323.
40. X.-H. Li, F. Yin, and C.-L. Liu, "Page segmentation using convolutional neural network and graphical model," in *Document Analysis Systems*, X. Bai, D. Karatzas, and D. Lopresti, Eds. Cham: Springer International Publishing, 2020, pp. 231–245.
41. R. Saha, A. Mondal, and C. V. Jawahar, "Graphical object detection in document images," *CoRR*, vol. abs/2008.10843, 2020. [Online]. Available: https://arxiv.org/abs/2008.10843
42. K. Li, C. Wigington, C. Tensmeyer, H. Zhao, N. Barmpalios, V. I. Morariu, V. Manjunatha, T. Sun, and Y. Fu, "Cross-domain document object detection: Benchmark suite and method," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 915–12 924.
43. H. Yang and W. H. Hsu, "Vision-based layout detection from scientific literature using recurrent convolutional neural networks," in *2020 25th international conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 6455–6462.
44. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
45. J. Li, Y. Xu, T. Lv, L. Cui, C. Zhang, and F. Wei, "Dit: Self-supervised pre-training for document image transformer," 2022. [Online]. Available: https://arxiv.org/abs/2203.02378
46. Y. Li, Y. Qian, Y. Yu, X. Qin, C. Zhang, Y. Liu, K. Yao, J. Han, J. Liu, and E. Ding, "Structext: Structured text understanding with multi-modal transformers," *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:236950714
47. R. Powalski, Ł. Borchmann, D. Jurkiewicz, T. Dwojak, M. Pietruszka, and G. Pałka, "Going full-tilt boogie on document understanding with text-image-layout transformer," in *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*. Springer, 2021, pp. 732–747.
48. H. Yang and W. Hsu, "Transformer-based approach for document layout understanding," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 4043–4047.
49. X. Zhong, J. Tang, and A. J. Yepes, "Publaynet: largest dataset ever for document layout analysis," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, Sep. 2019, pp. 1015–1022.
50. Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "Layoutlmv3: Pre-training for document ai with unified text and image masking," 2022. [Online]. Available: https://arxiv.org/abs/2204.08387

51. S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, and R. Manmatha, "Docformer: End-to-end transformer for document understanding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 993–1003.
52. G. Kim, T. Hong, M. Yim, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park, "Donut: Document understanding transformer without OCR," *CoRR*, vol. abs/2111.15664, 2021. [Online]. Available: https://arxiv.org/abs/2111.15664
53. J. Gu, J. Kuen, V. I. Morariu, H. Zhao, R. Jain, N. Barmpalios, A. Nenkova, and T. Sun, "Unidoc: Unified pretraining framework for document understanding," *Advances in Neural Information Processing Systems*, vol. 34, pp. 39–50, 2021.
54. Z. Gu, C. Meng, K. Wang, J. Lan, W. Wang, M. Gu, and L. Zhang, "Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4583–4592.
55. T. Shehzadi, K. A. Hashmi, D. Stricker, M. Liwicki, and M. Z. Afzal, "Bridging the performance gap between detr and r-cnn for graphical object detection in document images," *arXiv preprint arXiv:2306.13526*, 2023.
56. T. Shehzadi, K. A. Hashmi, D. Stricker, and M. Z. Afzal, "Sparse semi-detr: Sparse learnable queries for semi-supervised object detection," *arXiv preprint arXiv:2404.01819*, 2024.
57. K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.
58. D. Gunawan, C. A. Sembiring, and M. A. Budiman, "The implementation of cosine similarity to calculate text relevance between two documents," *Journal of Physics: Conference Series*, vol. 978, no. 1, p. 012120, mar 2018. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/978/1/012120
59. F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "Dn-detr: Accelerate detr training by introducing query denoising," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 619–13 627.
60. B. Smock, R. Pesala, and R. Abraham, "PubTables-1M: Towards comprehensive table extraction from unstructured documents," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 4634–4642.
61. T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312
62. Z. Zhong, J. Wang, H. Sun, K. Hu, E. Zhang, L. Sun, and Q. Huo, "A hybrid approach to document layout analysis for heterogeneous document images," in *Document Analysis and Recognition - ICDAR 2023*, G. A. Fink, R. Jain, K. Kise, and R. Zanibbi, Eds. Cham: Springer Nature Switzerland, 2023, pp. 189–206.
63. N. Sun, Y. Zhu, and X. Hu, "Faster r-cnn based table detection combining corner locating," *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1314–1319, 2019.
64. A. Bochkovskiy, C. Wang, and H. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *CoRR*, vol. abs/2004.10934, 2020. [Online]. Available: https://arxiv.org/abs/2004.10934
65. M. Minouei, M. R. Soheili, and D. Stricker, "Document layout analysis with an enhanced object detector," in *2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, 2021, pp. 1–5.

66. H. Bi, C. Xu, C. Shi, G. Liu, Y. Li, H. Zhang, and J. Qu, "Srrv: A novel document object detector based on spatial-related relation and vision," *IEEE Transactions on Multimedia*, vol. 25, pp. 3788–3798, 2023.
67. P. Zhang, C. Li, L. Qiao, Z. Cheng, S. Pu, Y. Niu, and F. Wu, "VSR: A unified framework for document layout analysis combining vision, semantics and relations," *CoRR*, vol. abs/2105.06220, 2021. [Online]. Available: https://arxiv.org/abs/2105.06220