Graphical Abstract

Mamba-YOLO: Multi-Level Adaptive Rectangular Convolution for Document Layout Analysis

Wenkang Ma, Mingzhe Cao, Jinyue Ma, Zhenyang Dong, Chaozhi Yang, Zongmin Li*

Highlights

Mamba-YOLO: Multi-Level Adaptive Rectangular Convolution for Document Layout Analysis

Wenkang Ma, Mingzhe Cao, Jinyue Ma, Zhenyang Dong, Chaozhi Yang, Zongmin Li*

- A Document Layout Analysis Dataset focuses on fine granularity.
- A novel and effective Document Layout Analysis network.
- Mamba structure is introduced into Document Layout Analysis for the first time.
- Our approach is competitive both in PeKi and open source datasets.

Mamba-YOLO: Multi-Level Adaptive Rectangular Convolution for Document Layout Analysis

Wenkang Ma, Mingzhe Cao, Jinyue Ma, Zhenyang Dong, Chaozhi Yang, Zongmin Li*

[a]*Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao, 266580, Shandong, China*

Abstract

Document Layout Analysis (DLA) is a crucial component in the field of document understanding and processing. However, the majority of existing public DLA datasets are predominantly in English and confined to a single domain, which limit their applicability in the diverse context of Chinese documents. To address this limitation, we introduce PeKi, a large-scale dataset for Chinese document layout analysis that encompasses multiple domains. PeKi comprises a comprehensive collection of documents from five distinct fields, available in both scanned and digitally native formats. We have meticulously annotated all titles and ensured that figure captions and table captions are included within the respective scopes of figures and tables. Building on this foundation, we propose a novel approach to document layout analysis, termed DocMY. This method innovatively integrates Mamba module into YOLOv9 framework, for the first time, with a multi-level adaptive rectangular convolution. This can effectively perceive different layout elements with varying aspect ratios and understand structured continuous elements in the document. Experiments on the proposed benchmarks (PubLayNet, CDLA, DocLayNet, D4LA) demonstrate that DocMY obtains competitive results. DocMY achieves a mAP of 94.8% on PeKi dataset while reducing model parameters by 15.3% compared to previous methods. Our code and dataset are available on `https://github.com/WenkMa/DocMY`.

Keywords: Document Layout Analysis, Mamaba, YOLO

1. Introduction

As a core component of document understanding, the importance of Document Layout Analysis has become increasingly prominent[1, 2, 3]. Over the past few years, DLA has seen significant advancements. However, the existing work[4, 5, 6] is still based on English documents and captures only coarse-grained document layout information. Currently, Deep Learning technology occupies a central position in the field of DLA and continues to drive progress in this domain. LayoutLMv3[4] and DiT[6] adopt a multimodal approach, introduce a self-supervised training mechanism, and sense document layout based on Transformer[7] module, which can effectively make up for the problem of using only a single mode of images, and can be fine-tuned for downstream tasks[8].

However, the current research faces three challenges. Firstly, there is a notable absence of comprehensive Chinese datasets that span multiple domains and have complex layouts. Existing studies predominantly rely on English-language datasets, and even when Chinese datasets are available, they are often limited to specific fields or contain only a small proportion of Chinese content, and document layout analysis is simple and lacks complexity. Secondly, current methods do not adequately address objects with varying aspect ratios. Uniformly sized convolution kernels are insufficient for effectively adapting to this task, leading to suboptimal performance. Lastly, for objects that require continuity detection, there is a significant issue of feature loss. This results in the failure to recognize continuous elements, such as document titles, which are crucial for accurate layout analysis.

In order to fill the gap in comprehensive Chinese datasets that span multiple domains, we propose and construct a novel Chinese document layout analysis dataset. Dataset encompasses document samples from various domains, including university textbooks, industry regulation reports, government reports, organizational notifications,

and scientific papers, which we have named PeKi dataset. Compared to previously available datasets[5, 9, 10, 11], PeKi annotate and refining all the "Title" and focused on the fine-grained information. At the same time, the Figure, Table and Formula areas cover the caption. For object detection, YOLO[12] is an outstanding algorithm known for its outstanding accuracy and speed on natural images. In order to better play the performance of YOLO, and adapt it to the document layout analysis task, we adopt the efficient design principles of VMamba[13] and integrate Mamba[14] design philosophy into the YOLOv9[12], thereby constructing a new architecture named Document-Mamba-YOLO (DocMY). We propose an optimized convolutional kernel design strategy aimed at extracting features using Multi-level Adaptive Rectangular Convolution (MARConv), especially for objects with varying aspect ratios. For feature loss in continuity detection, especially when dealing with sequential titles, we introduce Mamba into the YOLOv9 [12] to build a linear relationship to extract continuity, and construct Generalized ELAN State Space Model module. This module can effectively perceive different layout elements with understand structured continuous elements in the document.

Experiments on the proposed benchmarks (D4LA[5], CDLA[9], PubLayNet[10], DocLayNet[11]) demonstrate that DocMY obtains competitive results. Compared to the baseline, our design incorporating two modules led to performance improvements across all the aforementioned datasets. Specifically, on the CDLA dataset, every category saw an increase in performance, with an mAP improvement of 2.1%. Moreover, DocMY achieved an mAP of 96.5%, a Precision of 94.8%, and a Recall of 94.0% on the PeKi dataset, attaining SOTA performance among various comparative methods. Additionally, the model's parameters were reduced by 16.3% compared to previous approaches.

Main contributions of this paper are summarized as follows:

1. We propose PeKi, a dataset that span multiple domains focused on Chinese

document layout analysis. This dataset annotates all the titles and optimizes annotation of Table and Figure with their captions.

2. We propose a novel method DocMY, which achieves SOTA performence on PeKi, introducing Mamba[14] into the task of document layout analysis for the first time. We improve the design of the convolutional structure by introducing Multi-level Adaptive Rectangular Convolution.
3. Our model effectively mitigates domain transfer issues, significantly improving baseline performance on benchmark suites. On several existing open-source datasets D4LA[5],CDLA[9], PubLayNet[10] and DocLayNet[11], our model demonstrated Precision improvements of 5.2%, 2.2%, 1.0%, and 2.7%, respectively.

2. Related Work

2.1. Document Layout Analysis Datasets

Document Layout Analysis datasets such as PubLayNet[10] consists of 360,000 images primarily focusing on five types of document elements: titles, text, images, tables, and lists. Annotations are mainly auto generated. CDLA[9] dataset is a Chinese document layout analysis dataset tailored for scholarly publication, a single domain. DocLayNet[11] manually annotates four languages and six document types, comprising 11 categories of layout analysis from diverse document types. D4LA[5] select 11,092 images from the RVL-CDIP[15] dataset to form 27 categories, further adding detailed annotations for emails, resumes, etc. DocSynth-300K[16] is generated using a diffusion model[17]. Other datasets [18, 19, 20, 21] are either not open-source or are primarily suited for downstream task fine-tuning. Most of the datasets focus on English-language contexts. Overall, current document layout analysis datasets have significant limitations in language and diversity.

2.2. Document Layout Analysis Approaches

Document Layout Analysis focuses on identifying and locating different elements within documents, like "title" and "text". DLA methodologies can be broadly categorized into three approaches: heuristic rulebased designs[22], Machine Learning, and Deep Learning. Initially, DLA methods relied on heuristic rule designs by reading many domain-specific documents and designing specific heuristic rules[22]. Researchers are starting to adopt machine learning methods [4, 6, 23, 24, 25, 26] because traditional methods require labor costs and are poorly scalable. Faster R-CNN[27] considered that DLA can be treated as a specialized object detection problem. Recently, the advent of Transformers and multi-modal approaches has significantly advanced the field. DocLayout-YOLO [16] is better able to handle multi-scale variations of document elements. From the global page scale to the local semantic information, Global-to-Local Design-Controllable Receptive Module enables models to efficiently detect targets at different scales. DiT[6] has trained a document image Transformer for DLA, employing self-supervised learning from large-scale unlabeled document images to enhance performance. LayoutLM[24, 25, 4] integrate text, layout, and image for pre-training purposes, followed by fine-tuning on downstream tasks, achieving impressive results on various document tasks. But their parameters are too large and require additional tools to extract text and layout.

2.3. Vision Mamba

State Space Models (SSMs) have been a focal point of recent research. Building upon SSMs[28], the Mamba[14] study introduced linear complexity, addressing the computational efficiency issues of Transformers on long sequences. VMamba[13] and Vision Mamba[29] are pure vision backbone, marking the first introduction of Mamba into the visual domain. DocMamba[30] introduces Segment-First Bidirectional Scan to capture continuous semantic information, reducing memory usage. Vision Mamb[29]

and VMamba [13] still hold considerable potential for further exploration, particularly in the DLA.

3. PeKi Dataset

Dataset	#Image	#Class	#Instance	Annotating Means	Format	Document Type	Language
PubLayNet [10]	360K	5	3,311,660	Automatic	PDF	Articles	English
CDLA [9]	6K	10	70,928	Manual	PDF	Articles	Chinese
DocLayNet [11]	81K	11	1,107,470	Manual	PDF	Financial Reports, Manuals, Scientific Articles, Laws, Regulations, Patents, Government Tenders	English, German, French, Japanese
D4LA [5]	11K	27	146,846	Automatic	PDF	Ideological, Moral Cultivation, Basic Law Education	English
M6Doc [18]	9K	74	237,116	Manual	PDF, Scanned, Photographed	Scientific Articles, Textbooks, Books, Test papers, Magazines, Newspapers, Notes	English, Chinese
DocBank [20]	500K	13	-	Automatic	PDF	Articles	English
TableBank [21]	4K	1	417,234	Automatic	Latex, Word	Articles	English
DocSynth-300K [16]	11K	10	109,004	Automatic	PDF	Academic, Textbook, Market Analysis, Financial	Chinese
PeKi (**Ours**)	46K	14	373,656	Manual	Scanned, Word, PDF	University Books, Industry Regulation Reports, Government Reports, Unit Notices, Scientific Articles	Chinese

Table 1: Modern Document Layout Analysis Datasets.

Category	Training / Validate / Test	Category	Training / Validate / Test
Formula	285 / 19 / 46	Doc-Title	3451 / 420 / 392
Formula-Num	273 / 20 / 45	Footer	26498 / 3357 / 3307
Figure	7682 / 921 / 928	Header	13786 / 1728 / 1725
Figure-Caption	3926 / 515 / 496	Title-NoId	1648 / 245 / 202
Reference	730 / 78 / 93	Title-Id	214037 / 27951 / 26565
Table	16798 / 2052 / 2127	Title-Last	1060 / 151 / 117
Table-Caption	7375 / 929 / 930	Title-Body	587 / 79 / 82

Table 2: The number of different document layout types on PeKi.

PeKi collected a total of 50,121 document images from various online sources[1]. Following a preprocessing stage that involved filtering out excessively unclear or blank images, 46,777 high-quality images are retained for annotation. The dataset was subsequently annotated by a team comprising twelve graduate students and three researchers, resulting in a total of 373,656 labeled instances.

[1]Partial URLs: https://www.jcad.cn, https://www.zhuanzhi.ai/, https://www.mofcom.gov.cn/

The annotations created using the Labelme labeling software provide detailed layout information in the form of labeled rectangular bounding boxes. Specifically, we define 14 distinct layout categories: Doc-Title, Title-Id, Title-NoId, Title-Body, Title-Last, Formula, Formula-Num, Figure, Figure-Caption, Table, Table-Caption, Reference, Footer, and Header. A comprehensive description of these categories is provided in Appendix A.2.

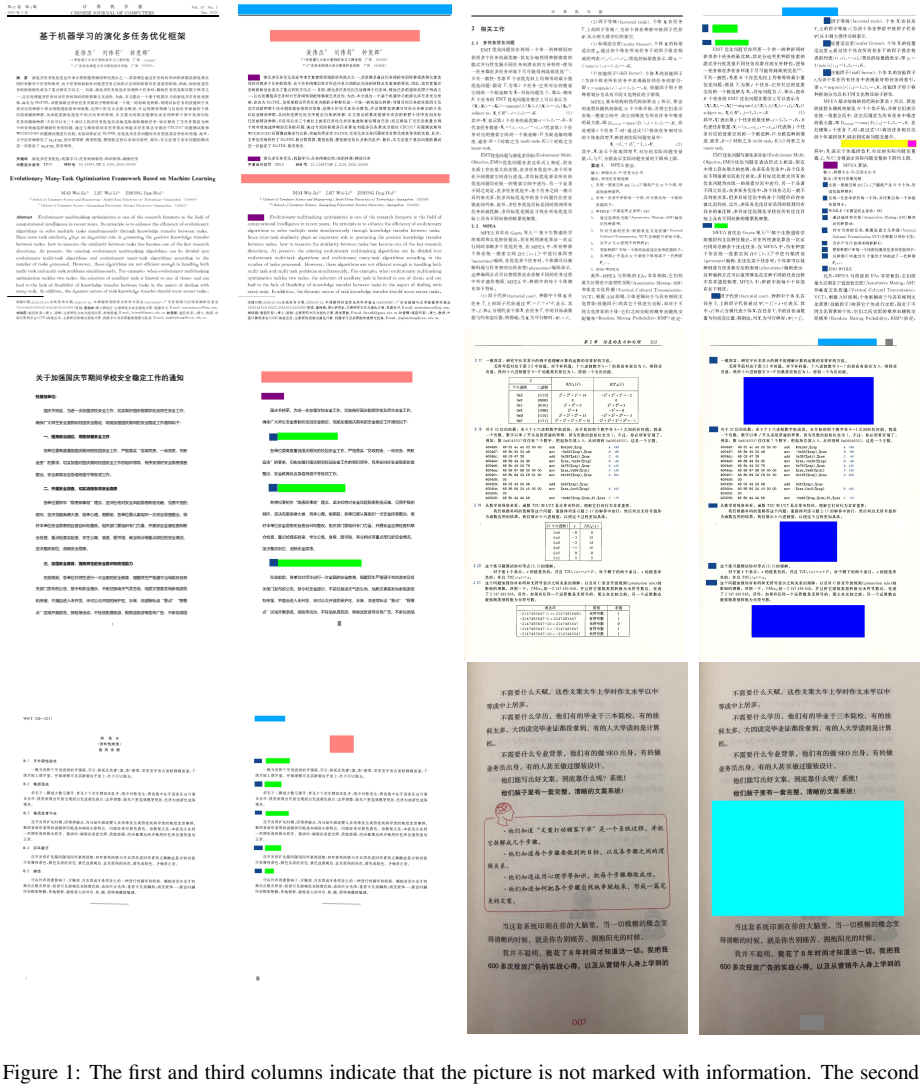

Figure 1: The first and third columns indicate that the picture is not marked with information. The second and fourth columns show the picture with visual annotation. Different colors indicate different categories.

Existing large-scale Document Layout Analysis datasets [9, 10] predominantly comprise scientific papers, which lack of display documents document layout analysis and fine-grained layout analysis. In contrast, PeKi encompasses a broader selection of five document types: computer science journals, university textbooks, industry standard reports, government reports, and organizational notifications. These document types are further categorized into 14 fine-grained layout classes, as illustrated in Figure 1 and detailed in Table 1, enhancing the dataset's applicability to real-world scenarios and the complexity of layout analysis.

A key contribution of PeKi is the introduction of fine-grained 'Title' annotations, which classify 'Title' into five distinct categories: Doc-Title, Title-Id, Title-NoId, Title-Body, and Title-Next. This refined segmentation increases the complexity of document layout analysis, requiring models to capture more nuanced structural information. From Figure 1 and Table 2, we focus on the fine-grained information of 'Title'[5, 18], such as third level and fourth level titles. It is evident that we refines the 'Title' attribute from previous work into more detailed segments. Specifically, we introduce the following categories: 'Doc-Title' represents the title information on the document cover or scientific article topic; 'Title-Id' indicates the serial number of the title and does not include any subsequent text information; 'Title-NoId' is used for titles that are in bold but do not have an title index; 'Title-Body' is for titles that are both indexed and in bold; 'Title-Last' is used when the 'Title-Body' is too long and the title information is wrapped to the last line. These refinements enable a more nuanced and accurate representation of title information, enhancing the overall quality and usability of the dataset. Like CDLA [9], we also pay attention to the fine-grained information of 'Table-Caption', 'Figure-Caption' and 'Formula-Num', captions annotation area needs to be within the scope of the Figure, Table or Formula annotation. Additionally, the dataset exhibits an irregular distribution of layout elements, introducing challenges in diversity and detection stability. Further complexities arise from real-world imperfec-

tions such as noise, blurriness, and other artifacts, making PeKi a more demanding benchmark for layout analysis.

The diversity and complexity of the PeKi dataset make it a valuable resource for advancing research in document layout analysis. To facilitate consistent benchmarking within the document layout analysis community, we have pre-defined train, validation, and test splits in an 8:1:1 ratio. This eliminates variations in evaluation scores caused by random dataset splitting. Additionally, we ensured that less frequent labels are evenly across the train and test sets, maintaining a balanced distribution for robust model evaluation. The distribution of annotation instances across categories is presented in Table 2.

PeKi comprises five distinct document types, each presenting unique challenges for document layout analysis:

University Textbooks: University textbooks often contain a diverse range of elements, including paragraph, mathematical formulas, pseudocode blocks, and illustrative figures and tables. These heterogeneous components frequently coexist in irregularly nested structures, further complicated by the presence of numerous multi-level titles.

Industry Regulation Reports: Industry regulation reports demand meticulous attention to document details, particularly in the fine-grained classification of title information. The rigid formatting standards and hierarchical title structures at varying granular levels introduce significant challenges for document layout analysis.

Government Reports: Government reports often use color and bold text to differentiate content and refine document structure. Additionally, they incorporate labeled title to organize information. However, the distribution of these titles is inconsistent, and their positions within the document are unpredictable, posing challenges for structured layout analysis.

Organizational Notifications: Organizational notifications follow specific docu-

ment formatting conventions; however, variations across different organizations lead to inconsistencies in content structure. Additionally, these documents often contain a significant amount of untitled content 'Title-Id', requiring more human judgment and discussion during the annotation process to ensure accurate labeling.

Scientific Articles: Scientific articles, just like CDLA [9], primarily require layout analysis of figures, tables, and mathematical formulas while also necessitating fine-grained segmentation of document titles to accurately capture their hierarchical structure.

We opted for manual annotation in the PeKi dataset instead of automated methods due to several key advantages. Manual annotation enables labeling across diverse document types without requiring programming expertise. Additionally, many documents lack original digital versions, rendering automated annotation impractical. Human annotation ensures higher accuracy and provides a more natural interpretation of page layouts. For instance, when identifying a 'Title-Id', human annotators can assess whether a text element genuinely functions as a title, whereas automated methods often rely on metadata that may be incomplete or inaccurate. Similarly, in table annotations, certain documents contain tables embedded as images rather than structured data. Automated systems may misclassify these as 'Figure', introducing errors that human annotators can avoid. To ensure annotation quality, each image undergoes multiple rounds of review and validation. The process includes an initial annotation phase, verification by additional annotators, visualization checks, and final validation by research experts. This rigorous quality control framework enhances data accuracy and ensures a high-quality dataset, making PeKi a valuable benchmark for document layout analysis, particularly in fine-grained structural understanding across diverse document types.

The PeKi dataset sets itself apart from existing document layout analysis datasets, such as D4LA [5], CDLA [9], PubLayNet [10], and DocLayNet [11], through the

following distinctive characteristics:

1. Refined Title Annotations: PeKi provides a more granular segmentation of 'Title', distinguishing between different types to enhance document structure analysis.
2. Precise Caption and Numbering Annotations: Unlike previous datasets, PeKi ensures that 'Figure-Caption', 'Table-Caption', and 'Formula-Num' are annotated strictly within their respective object boundaries, improving spatial accuracy.
3. Exclusion of the 'Text' Category: Unlike prior works that include a generic Text category, PeKi discards this label, focusing instead on more meaningful structural components. However, we still make annotations for the text content of the area of interest, such as 'caption','Title-Body' and 'Title-Last'.
4. Selective Title Index Annotation: In sentences prefixed by a title, only the title index is annotated, without including the corresponding textual content, maintaining a clear distinction between structural and semantic elements.

4. Method

4.1. Preliminaries

Structured state-space sequence models, such as S4 [28] and Mamba [14], are derived from state-space models and operate on a continuous dynamical system that maps an input sequence $x(t) \in \mathbb{R}$ to an output sequence $y(t) \in \mathbb{R}$ via an implicit latent state $h(t) \in \mathbb{R}^N$. This formulation effectively captures temporal dependencies and models sequential data dynamics. The continuous system is defined as follows:

$$\mathbf{h}'(t) = \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{x}(t), \tag{1}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{h}(t) + \mathbf{D}\mathbf{x}(t). \tag{2}$$

In Equation (1), $\mathbf{A} \in \mathbb{R}^{N\times N}$ represents the state transition matrix, governing the evolution of the hidden state over time, while $\mathbf{B} \in \mathbb{R}^{N\times 1}$ is the input transformation matrix. In Equation (2), $\mathbf{C} \in \mathbb{R}^{N\times 1}$ is the observation matrix, responsible for mapping the hidden state to the output, and $\mathbf{D} \in \mathbb{R}^{N\times 1}$ is the input-output skip-connection matrix.

To adapt this continuous system for discrete-time sequence modeling, Mamba applies fixed discretization rules, transforming $\mathbf{A}$ and $\mathbf{B}$ into their discrete counterparts $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$, respectively. A commonly used discretization method for this transformation is the Zero-Order Hold (ZOH), which is expressed as:

$$\overline{\mathbf{A}} = \exp(\mathbf{\Delta A}), \tag{3}$$

$$\overline{\mathbf{B}} = (\mathbf{\Delta A})^{-1}(\exp(\mathbf{\Delta A}) - \mathbf{I})\mathbf{\Delta B}. \tag{4}$$

Here, $\mathbf{\Delta}$ represents a time-scale parameter that modulates the temporal resolution of the model, while $\mathbf{\Delta A}$ and $\mathbf{\Delta B}$ are the discrete-time equivalents of their continuous counterparts. The identity matrix is denoted by $\mathbf{I}$. After discretization, the model operates in a linear recursive form:

$$h'(t) = \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t, \tag{5}$$

$$y_t = \mathbf{C}h_t + \mathbf{D}x_t. \tag{6}$$

Additionally, the entire sequence transformation can be reformulated as a convolution, expressed as:

$$\overline{\mathbf{K}} = (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}\mathbf{B}}, \ldots, \mathbf{C}\overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}}), \tag{7}$$

$$y = x * \overline{\mathbf{K}} + \mathbf{D}x_t. \tag{8}$$

Here, $\overline{\mathbf{K}} \in \mathbb{R}^L$ represents the structured convolutional kernel, where L denotes the length of the input sequence. In the proposed approach, the model leverages the convo-

lutional formulation for parallel training while utilizing the linear recursive representation for efficient autoregressive inference. This design enables the model to capture long-range dependencies and effectively model sequential data dynamics.

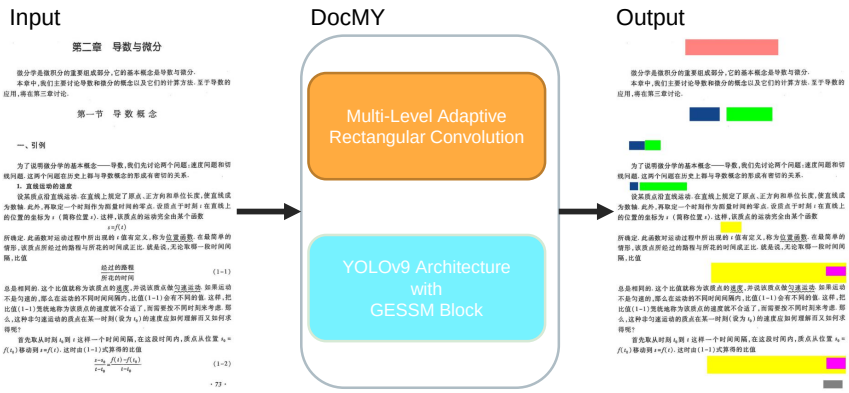

Figure 2: DocMY Model Architure

4.2. Model Architecture

Document layout analysis aims to identify and annotate all relevant objects within a document image. While spatial feature pyramid methods effectively detect small objects, they often fail to consider variations in object sizes and lack sensitivity to contextual information within the image. To address the challenges posed by inconsistent aspect ratios of objects and the need for capturing document-specific contextual features, we propose a novel method named **DocMY**. The overall architecture of DocMY is illustrated in Figure 2 and Figure 3, where a Generalized ELAN State Space Model (GESSM) Block is integrated into the YOLOv9 [12] framework with a Multi-level Adaptive Rectangular Convolution Feature Extraction (MARConv) module. The MARConv module is designed to capture multi-scale object features by employing convolutional kernels of three different sizes, mitigating the limitations of fixed-size

kernels when processing objects of varying dimensions. A detailed description of this module is provided in Section 4.3.

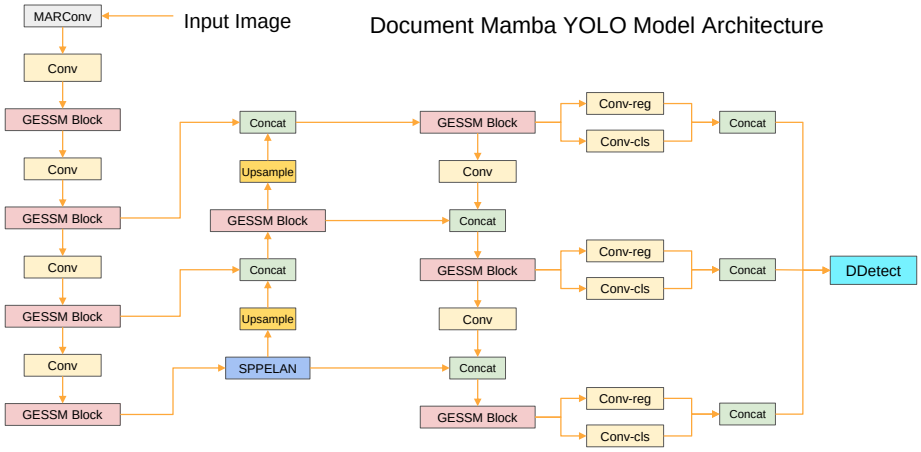

Figure 3: Document Mamba YOLO Model Architure

After extracting fine-grained document features, a multi-scale architecture is employed to capture long-range contextual dependencies within the document image. To further enhance contextual feature extraction, we introduce the GESSM Block, which utilizes a scanning mechanism to extract structural information from document images. Additionally, the embedded State Space Model module enables the extraction of deep hierarchical features. The final feature representation is obtained through feature fusion, ensuring an optimal representation of the document layout. A comprehensive explanation of the GESSM Block is provided in Section 4.4.

4.3. Multi-level Adaptive Rectangular Convolution Feature Extraction

The Spatial Feature Pyramid is a multi-scale feature fusion technique that uses convolution kernels of different sizes (1×1, 3×3, 5×5) at different layers to capture objects of varying granularities, making it highly effective in object detection. However, in document layout analysis, it struggles with structural adaptability and fails to

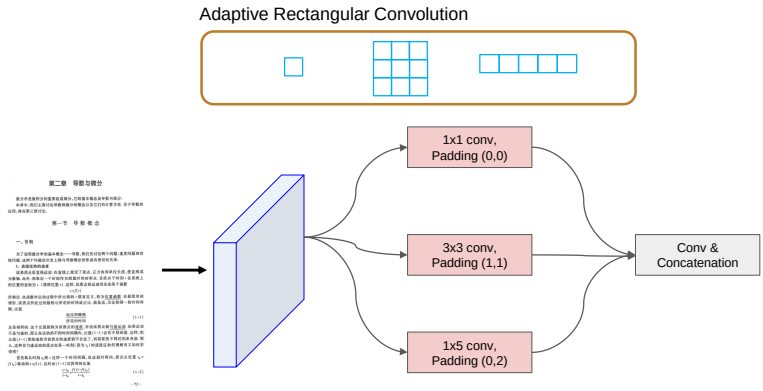

Figure 4: Multi-Level Adaptive Rectangular Convolution

capture fine-grained details due to the large aspect ratios of objects. Inspired by deformable and adaptive convolutions [31, 32, 33] and spatial feature pyramids [34, 35], we propose Multi-Level Adaptive Rectangular Convolution, a three-level convolutional module designed to effectively extract multi-scale features while addressing the limitations of fixed-size convolution kernels in handling objects of varying dimensions.

In our approach, we modify the original 5×5 convolution kernel in the spatial feature pyramid to 1×5 while keeping other kernel structures unchanged, as shown in Figure 4. The design of rectangular convolutional kernels can better adapt to object detection tasks with significant aspect ratio variations, improving the detection of large-scale objects while effectively capturing small object features [33, 32]. Leveraging this advantage, we incorporate it into our proposed multi-level adaptive feature extraction module to enhance the model's representation capability for objects of different scales. By extracting image features at different scales in a hierarchical manner, we mitigate feature dimension mismatches caused by varying kernel sizes.

This method leverages the advantages of spatial feature pyramids for capturing hierarchical object features while enhancing adaptability for document layout analysis.

The core mechanism is formulated as:

$$
\begin{aligned}
Y^{(1\times1)} &= \sigma\left(\mathrm{Conv}_{1\times1}(X; K_{1\times1}, p = (0,0)) + b_1\right), \\
Y^{(3\times3)} &= \sigma\left(\mathrm{Conv}_{3\times3}(X; K_{3\times3}, p = (1,1)) + b_3\right), \\
Y^{(1\times5)} &= \sigma\left(\mathrm{Conv}_{1\times5}(X; K_{1\times5}, p = (0,2)) + b_2\right), \\
O &= Y^{(1\times1)} + Y^{(3\times3)} + Y^{(1\times5)},
\end{aligned}
\tag{9}
$$

where X denotes the input feature map $\mathbb{R}^{H\times W\times C}$, $Y^{(k)}$ represents the feature map extracted by the k-th convolutional layer, K_k denotes the kernel of size k, p denotes the padding size, and b denotes the bias term, σ denotes the activation function. The output O is the sum of the three feature maps, which are concatenated to form the final feature map. This design enables the model to adapt to varying aspect ratios and capture features at different scales, thereby enhancing the model's performance in document layout analysis.

4.4. Generalized ELAN State Space Model

The YOLO series is known for its efficient architecture and multi-scale module design, offering real-time performance and lightweight characteristi cs. However, in document layout analysis, it struggles to effectively capture rich contextual information. To retain YOLO's original efficiency while adapting it to document layout analysis and mitigating feature loss, we incorporate the Mamba module into YOLOv9 and refine the model architecture. This modification helps maintain competitive results while reducing both model parameters and computational complexity. YOLOv9 [12] introduces the Generalized ELAN (GELAN) network structure, which balances lightweight design, inference speed, and accuracy by leveraging ELAN for efficient feature aggregation. Without compromising its performance, we integrate the core SSM module from Mamba into GELAN, introducing the GESSM block. The structure of the GESSM block is illustrated in Figure 5 and Figure 6.

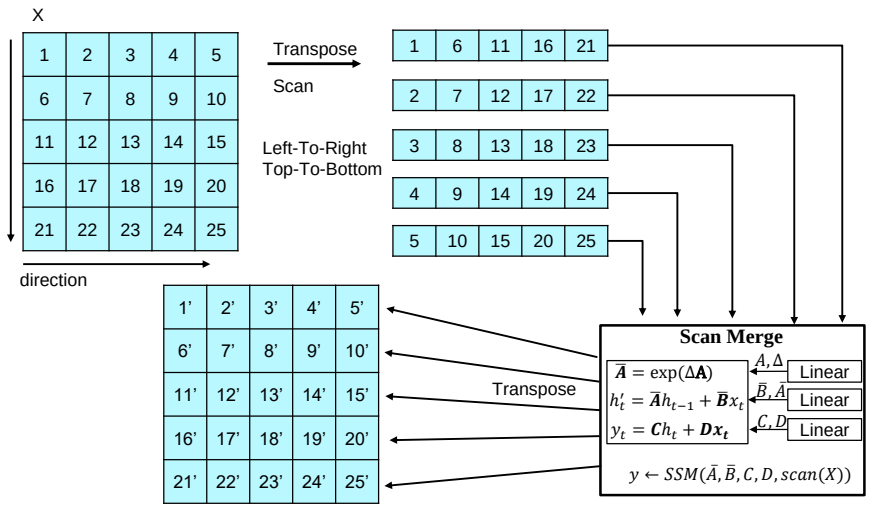

Figure 5: Scanning mechanism and structured state space processing flow

The first step involves image feature scanning to extract linear feature information. VMamba [13] introduces cross-scan module to solve orientation sensitive problem, but it is not suitable for document layout analysis. We modify the scanning mechanism by shifting the scanning process earlier in the pipeline. Since document layout analysis requires an awareness of positional structures within an image, this adjustment helps the model better capture relevant features from different layout regions, each associated with a specific directional scan. The process follows a left-to-right, top-to-bottom sequence, dynamically adjusting the scanning direction, as illustrated in Figure 5. This layout ensures full coverage of the input image while leveraging systematic vertical scanning to efficiently capture continuous spatial information. The result is a rich multi-dimensional feature library that enhances the efficiency and comprehensiveness of multi-dimensional feature extraction.

Another key aspect of the GESSM module is the integration of the State Space Model module, while still adhering to the GELAN architecture, as shown in Figure 6. After linear feature extraction, the state representation and weight matrices A, B, and C

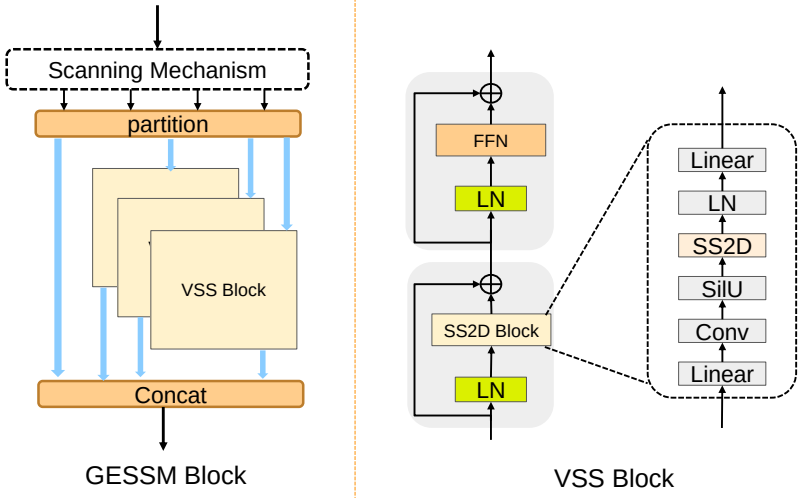

Figure 6: Generalized ELAN State Space Model

are used to extract deep feature information from the image, as shown in Equation 10.

$$\begin{aligned} \mathbf{h}'(t) &= \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{x}(t), \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{h}(t) + \mathbf{D}\mathbf{x}(t). \end{aligned} \tag{10}$$

The adjusted scanning strategy further enhances feature extraction efficiency for critical regions within the document layout. To address feature loss, we introduce a residual connection strategy, fusing the original features with those extracted through SSM. This mitigates feature degradation caused by state-space sequence modeling, as shown in Equation 11.

$$\begin{aligned} \text{SS2D Block}(X) &= \text{Linear}(LN(\text{SS2D}(SiLU(Conv(Linear(X)))))) \\ X' &= \text{SS2D Block}(\text{LN(X)}) \bigoplus X \\ \text{layer}_i &= FFN(LN(X)) \bigoplus X' \end{aligned} \tag{11}$$

Finally, all extracted features are merged through a feature fusion process, as shown in Equation 12. By leveraging Mamba's sequence modeling capabilities, our approach

enhances YOLO's global contextual understanding, significantly improving its text line detection performance in complex document layout analysis tasks.

$$\text{GESSM Block}_{o}utput = \text{concat}(\text{layer}_0, \text{layer}_1, \text{layer}_2, X) \tag{12}$$

5. Experiments

5.1. Experiment Metrics and Datasets

For comprehensive performance evaluation, we adopt the COCO-style evaluation protocol [36] with the following metrics: Precision, Recall, mAP@50, mAP@50:95, and model parameter count. The mathematical formulations are defined as follows:

$$\begin{aligned}
\text{Precision} &= \frac{|\text{TP}|}{|\text{TP}| + |\text{FP}|} \\
\text{Recall} &= \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}|} \\
\text{AP} &= \frac{1}{11} \sum_{r \in \{0, 0.1, \ldots, 1.0\}} \max_{\tilde{p} \geq r} \text{Precision}(\tilde{p}) \\
\text{mAP@50} &= \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i|_{IoU=0.5} \\
\text{mAP@50:95} &= \frac{1}{N} \sum_{i=1}^{N} \frac{1}{10} \sum_{t=0.5}^{0.95} \text{AP}_i|_{IoU=t}
\end{aligned} \tag{13}$$

Where:

- **True Positives (TP)**: Correct predictions with matched class label and IoU ≥ threshold
- **False Positives (FP)**: Incorrect predictions including class mismatches or IoU < threshold
- **False Negatives (FN)**: Undetected ground truth instances
- **Average Precision (AP)**: Area under precision-recall curve for individual class

	Method	P(%)	R(%)	mAP50(%)	mAP50-95(%)	parameters
Multimodal	Layoutlmv3[4]	92.8	73.9	94.2	-	133,267,834
	DiT[6]	89.7	86.4	91.4	84.3	87,679,409
Unimodal	YOLOv5[37]	91.6	**94.3**	95.6	77.2	21,775,401
	YOLOv7[38]	92.2	93.1	96.0	81.3	22,194,944
	YOLOv8[39]	93.3	76.5	85.6	70.1	25,907,988
	YOLOv9[12]	91.6	89.9	94.1	81.9	25,723,688
	YOLOv10[40]	88.2	84.9	90.2	78.4	**20**, **429**, **656**
Mamba	VMamba[13]	84.7	81.1	87.2	69.4	58,303,956
	Vision Mamba[29]	84.3	89.6	92.1	77.4	26,799,380
Mamba-YOLO	DocMY(ours)	**94.8**(+**3.2**)	94.0(+4.1)	**96.5**(+**1.4**)	**83.3**(+**1.4**)	21,799,380

Table 3: Experimental results of PeKi. **Bold** indicates the best results on the PeKi dataset.(+3.2) represents a performance improvement over baseline YOLOv9.

- **mAP@50**: Mean AP across all classes at IoU=0.5
- **mAP@50:95**: Mean AP averaged over 10 IoU thresholds (0.50-0.95, step=0.05)
- N: number of classes

To validate the effectiveness of DocMY, we conduct extensive experiments on the PeKi dataset and four public document layout analysis benchmarks, including D4LA[5], CDLA[9], PubLayNet[10] and DocLayNet[11].

5.2. Experiment Details

The hardware configuration for our experiments is as follows: an Intel(R) Xeon(R) Silver 4216 CPU @ 2.10GHz, 128GB of system memory, and four NVIDIA GeForce RTX 2080 Ti GPUs, each equipped with 11GB of VRAM. All models in this study were trained from scratch over a total of 100 epochs. The initial learning rate was set to 1×10^{-2} and gradually reduced to 1×10^{-5}. A linear warm-up strategy was applied during the first three epochs, after which the learning rate decay was adjusted based on model scale. Additionally, mosaic data augmentation was disabled for the final 15 epochs to stabilize training.

5.3. Results on PeKi Datasets

As a Chinese dataset, PeKi not only fills the gap in the current Chinese-focused document layout analysis, but also makes the classification of 'Title' more detailed and

the layout analysis more complex and diverse. The uncertainty of the location of different types of titles brings great challenges to layout analysis. To further verify the scientific and challenging nature of our proposed dataset, we conducted experiments in three different types of models. Experiments are conducted comparing Multimodal Transformer-based architectures such as LayoutLMv3[4] and DiT[6] and Unimodal YOLO, as well as VMamba[13] and Vision Mamba[29]. Experimental results of different methods on our dataset are presented in Table 3.

Experimental results on the PeKi dataset demonstrate significant performance differences between multimodal and unimodal approaches in document layout analysis. Among multimodal models, LayoutLMv3 achieved a detection precision of 92.8% and a recall of 73.9%, while DiT attained a detection precision of 89.7% and a recall of 86.4%. These methods are effective in verifying detected objects but struggle to comprehensively identify all targets. Additionally, due to the inherent complexity of Transformer-based architectures, these models have a large number of parameters, leading to increased computational costs during training.

In contrast, unimodal methods such as the YOLO series, which employ multi-scale structural designs, offer a more lightweight and efficient training process. YOLOv9 achieved a detection precision of 91.6%, a recall of 89.9%, mAP@50 of 94.1%, and mAP@50-95 of 81.9%. While these methods effectively capture object information in document images, they exhibit limitations in detecting elements such as 'Header' and 'Footer' and perform suboptimally when identifying large structural components such as 'Table' and 'Figure'.

The Mamba series, which incorporates the SSM module to leverage linear relationships and scanning mechanisms, demonstrated improvements in capturing document continuity. VMamba achieved a detection precision of 84.7%, a recall of 81.1%, mAP@50 of 87.2%, and mAP@50-95 of 69.4%, while Vision Mamba attained a detection precision of 84.3%, a recall of 89.6%, mAP@50 of 92.1%, and mAP@50-95

Class	Box(P)	R	mAP50	mAP50-95
All	0.948	0.940	0.965	0.833
Footer	0.983	0.864	0.974	0.687
Header	0.986	0.973	0.988	0.923
Reference	0.917	1.000	0.995	0.979
Doc-Title	1.000	0.954	0.995	0.844
Title-Id	0.993	0.939	0.992	0.803
Figure	0.965	0.927	0.965	0.946
Figure-Caption	0.930	0.913	0.933	0.864
Table	0.972	0.980	0.989	0.959
Table-Caption	0.911	1.000	0.984	0.920
Formula	0.797	0.786	0.771	0.590
Formula-Caption	1.000	0.941	0.995	0.588
Title-NoId	0.965	0.961	0.989	0.790
Title-Body	0.941	0.919	0.940	0.876
Title-Last	0.912	1.000	0.995	0.895

Table 4: Detailed experimental results per class category

of 77.4%. Although these models exhibit notable gaps in detection precision and recall, their ability to effectively capture document continuity mitigates issues related to header and footer detection while improving the recognition of large structural elements.

Our proposed method demonstrates superior performance on the PeKi dataset, achieving a detection precision of 94.8%, a recall of 94.0%, mAP@50 of 96.5%, and mAP@50-95 of 83.3%. Compared to the baseline YOLOv9's [12], our approach improves precision by 3.2%, recall by 4.1%, mAP@50 by 1.4%, and mAP@50-95 by 1.4%. Additionally, our method utilizes only 21,799,380 parameters, significantly fewer than YOLOv9's 25,723,688 parameters, while achieving a substantial precision improvement. This represents only 15.3% of YOLOv9 parameter count, validating the efficiency of our approach. Moreover, our method requires just 16.3% of the parameters used by LayoutLMv3 while surpassing its performance, further demonstrating its effectiveness.

To further evaluate the effectiveness of the DocMY method on the proposed dataset,

we analyzed its detection performance across all categories, as detailed in Table 4. DocMY achieves a perfect recall of 1.00 for Table Caption, Reference, and Title-Last, indicating that the model effectively captures and accurately detects these elements. Additionally, it achieves a high recall of 0.980 for Tables, demonstrating its strong capability in identifying structured tabular data. For Doc-Title and Formula Caption, the model achieves 100% precision, correctly identifying and classifying these elements with precision. Moreover, DocMY maintains strong detection performance across other categories, showcasing its robustness in document layout analysis. However, its performance in the Figure category is relatively weaker, despite the similarities between figures and tables. Visual analysis of intermediate steps reveals that the model struggles to differentiate between subfigures and full figures, often detecting both, which negatively impacts detection precision. Furthermore, the model exhibits suboptimal performance in detecting Formulas due to the inherent complexity of formula structures. The dataset includes a diverse range of mathematical expressions, such as handwritten formulas, multi-line equations in textbooks, and specialized formats like chemical and mechanical equations. The model's limited ability to learn these intricate recognition patterns leads to a decline in detection precision within this category.

5.4. Results on Other Datasets

To validate the effectiveness of the proposed DocMY method, we conducted experimental comparisons across multiple datasets in various domains with document layout analysis challenges. Table 5 presents the experimental results of our network architecture on datasets such as D4LA[5], CDLA[9], PubLayNet[10] and DocLayNet[11]. As shown in Table 5, the proposed model demonstrates varying degrees of improvement over the baseline. On the PubLayNet[10] dataset, although the state-of-the-art model achieves 96.2%, our model produces highly competitive results with fewer parameters. For the DocLayNet[11] dataset, recognition precision improves by 1.0%, and mAP50-95 performance increases by 1.1%. On the D4LA [5] dataset, the model

Datasets	Method	P(%)	R(%)	mAP50(%)	mAP50-95(%)
PubLayNet[10]	SOTA	-	-	-	96.2
	YOLOv9	87.2	86.5	88.1	83.1
	Ours	92.4(+**5.2**)	91.5(+**5.0**)	93.3(+**5.2**)	88.4(+**5.3**)
CDLA[9]	YOLOv9	90.1	87.4	94.0	77.3
	Ours	93.2(+**2.2**)	91.4(+**4.0**)	96.1(+**2.1**)	83.3(+**6.0**)
DocLayNet[11]	SOTA	-	-	-	76.8
	YOLOv9	88.5	81.8	89.6	69.8
	Ours	89.5(+**1.0**)	81.8	90.2(+**0.6**)	70.9(+**1.1**)
D4LA[5]	SOTA	-	-	-	68.8
	YOLOv9	75.1	64.1	69.8	56.0
	Ours	77.8(+**2.7**)	71.7(+**7.6**)	76.7(+**6.9**)	62.8(+**6.8**)

Table 5: Effects of Our Method on PubLayNet, CDLA, DocLayNet, and D4LA. **Bold** indicates performance improvement compared to baseline.red indicates SOTA performance. - indicates that we did not find or reproduce the result.

Method	Header	Text	Reference	Figure caption	Figure	Table caption
YOLOv9	92.4	96.9	96.4	92.3	97.0	95.1
Ours	**93.8**(+**1.4**)	**98.0**(+**1.1**)	**97.3**(+**0.9**)	**94.7**(+**2.4**)	**97.7**(+**0.7**)	**98.1**(+**3.0**)
Method	Table	Title	Footer	Equation	mAP	
YOLOv9	98.9	95.1	83.9	92.0	94.0	
Ours	**99.4**(+**0.5**)	**97.2**(+**1.9**)	**88.2**(+**4.3**)	**96.7**(+**4.7**)	**96.1**(+**2.1**)	

Table 6: Performance comparisons on CDLA dataset.

shows significant performance improvement, with a 2.7% increase in precision and a 7.6% increase in recall. Additionally, the mAP50-95 performance increases by 6.8%. However, there is still room for improvement between our proposed model and the best published results, but in terms of model lightweight, our model still has advantages.

5.5. Results on CDLA

Table 6 demonstrates that DocMY achieves competitive performance compared to the YOLOv9 model across all categories in the Chinese CDLA dataset[9]. Specifically, the model achieves a significant improvement in the detection of Table, Table Caption, and Footer, with an increase of 0.5%, 3.0%, and 4.3%, respectively. The model also achieves a 1.3% improvement in the detection of Header, and a 2.4% improvement in the detection of Figure Caption. The model's performance in detecting Text, Reference, Figure, and Title also improves by 1.1%, 0.9%, 0.7%, and 2.1%, respectively. The model's performance in detecting Equation improves by 4.7%, demonstrating the

Method	P(%)	R(%)	mAP50(%)	mAP50-95(%)
Baseline[12]	91.6	89.9	94.1	81.9
B+GESSM	93.5(+1.9)	93.9(+4.0)	96.9(+2.8)	81.9(+0.0)
B+GESSM+MARConv	**94.8**(+**3.2**)	**94.0**(+**4.1**)	**96.5**(+**2.4**)	**83.3**(+**1.4**)

Table 7: Ablation study of model Architure.

model's ability to detect complex mathematical formulas. The model's overall mAP also improves by 2.1%. We believe that one of the reasons for the model's strong performance on the CDLA dataset is the design of different convolution kernels to extract image features, which are then merged to capture target information at different scales. Additionally, the design of the GESSM module, which incorporates a long-context structure, enables the model to efficiently capture contextual information within the document images.

5.6. Ablation Experiment

We conduct thorough experiments to validate the effectiveness of GESSM Block and MARConv. Tab. 7 summarizes the ablation experiments conducted on our proposed PeKi dataset.

From Table 7, it can be observed that the baseline model (YOLOv9) achieves a precision of 91.6%, recall of 89.9%, mAP50 of 94.1%, and mAP50-95 of 81.9%. The baseline model, YOLOv9 [12], through its structural design, can extract image features and detect target information within documents. However, it fails to efficiently capture the contextual information within the image features and does not handle object detection across different scales effectively. After introducing the GESSM module, which utilizes scanning strategies and state-space transformations, the model can efficiently capture deep feature information in images. This leads to a significant performance improvement, with precision reaching 93.5%, recall improving to 93.9%, mAP50 increasing to 96.9%, and mAP50-95 achieving 81.9%. Compared to YOLOv9, this represents a 1.09% improvement in precision, a 4.0% improvement in recall, and a 2.8% increase in mAP50. These results demonstrate that the introduction of the GESSM

module enables DocMY to effectively capture contextual information within the image features. Furthermore, after incorporating the MARConv module, the model's performance improves even further. The precision reaches 94.8%, recall rises to 94.0%, mAP50 reaches 96.5%, and mAP50-95 increases to 83.3%. mAP50 improves by 1.3%, and mAP50-95 sees an increase of 1.4%. This demonstrates that the MARConv module further enhances the extraction of image features. The multi-scale structural design efficiently captures features of target objects at different scales in document images, while maintaining the model's stability and improving detection precision.

In summary, this ablation study demonstrates that incrementally introducing the State Space Model and the MARConv module effectively improves the overall performance of DLA, providing valuable insights for further optimization and the construction of efficient, robust document layout analysis.

6. Conclusion

In this paper, we introduce PeKi, a novel dataset composed of five types of PDF documents, including both digital and scanned versions, spanning 14 different categories. To our knowledge, PeKi is the first dataset to refine all title classifications and associate Figure, Table with captions, Formula, and serial numbers. This dataset provides a rich and diverse resources for document layout analysis. Based on PeKi dataset and several open source document layout analysis datasets, we propose DocMY, a novel document layout analysis model that integrates a state space model and multi-level adaptive rectangular convolution. The experiments conducted on extensive downstream datasets demonstrate that DocMY significantly outperforms existing methods on both the PeKi dataset and several public datasets. This underscores the effectiveness and generalizability of the DocMY model.

The strengths of our work are multifaceted. PeKi's unique feature lies in its fine-grained annotations, others can parse documents and build hierarchical relationships

from them based on title numbers and associations of figures, tables, and formulas with content in the body. DocMY leverages the SSM and MARConv to address limitations in traditional approaches to complex document layouts. However, this work still has some limitations. First, the PeKi dataset primarily covers Chinese documents, and there is room for improvement by incorporating more document types and refining label granularity. Second, the current approach struggles to effectively analyze extremely large objects. In the future, we plan to expand the PeKi dataset to include a wider variety of document types and languages, further increasing the complexity of document layout analysis. Additionally, incorporating multimodal information to enrich the input features of document layout analysis is also an interesting direction to explore.

Acknowledgments

This work is partly supported by National key r&d program (Grant no. 2019YFF0 301800), National Natural Science Foundation of China (Grant no. 61379106),the Shandong Provincial Natural Science Foundation (Grant nos. ZR2015FM011).

References

[1] L. Qiao, C. Li, Z. Cheng, Y. Xu, Y. Niu, X. Li, Reading order detection in visually-rich documents with multi-modal layout-aware relation prediction, Pattern Recognition 150 (2024) 110314. doi:https://doi.org/10.1016/j.patcog.2024.110314.

[2] J. Wang, K. Hu, Z. Zhong, L. Sun, Q. Huo, Detect-order-construct: A tree construction based approach for hierarchical document structure analysis, Pattern Recognition 156 (2024) 110836. doi:https://doi.org/10.1016/j.patcog.2024.110836.

[3] N. Raman, S. Shah, M. Veloso, Synthetic document generator for annotation-free layout recognition, Pattern Recognition 128 (2022) 108660. doi:https://doi.org/10.1016/j.patcog.2022.108660.

[4] Y. Huang, T. Lv, L. Cui, Y. Lu, F. Wei, Layoutlmv3: Pre-training for document ai with unified text and image masking, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 4083–4091. doi:10.1145/3503161.3548112.

[5] C. Da, C. Luo, Q. Zheng, C. Yao, Vision grid transformer for document layout analysis, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19405–19415. doi:10.1109/ICCV51070.2023.01783.

[6] J. Li, Y. Xu, T. Lv, L. Cui, C. Zhang, F. Wei, Dit: Self-supervised pre-training for document image transformer, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 3530–3539. doi:10.1145/3503161.3547911.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 6000–6010.

[8] L. Cui, Y. Xu, T. Lv, F. Wei, Document ai: Benchmarks, models and applications (2021). arXiv:2111.08609.

[9] https://github.com/buptlihang/cdla (2021).

[10] X. Zhong, J. Tang, A. Jimeno Yepes, Publaynet: Largest dataset ever for document layout analysis, in: 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1015–1022. doi:10.1109/ICDAR.2019.00166.

[11] B. Pfitzmann, C. Auer, M. Dolfi, A. S. Nassar, 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. doi:10.1145/3534678.3539043.

[12] C.-Y. Wang, I.-H. Yeh, H.-Y. Mark Liao, Yolov9: Learning what you want to learn using programmable gradient information, in: Proceedings of the International Conference on Computer Vision, 2025, pp. 1–21. doi:10.1007/978-3-031-72751-1_1.

[13] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, Y. Liu, Vmamba: Visual state space model (2024). arXiv:2401.10166.

[14] A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces (2024). arXiv:2312.00752.

[15] A. W. Harley, A. Ufkes, K. G. Derpanis, Evaluation of deep convolutional nets for document image classification and retrieval, in: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 2015, pp. 991–995. doi:10.1109/ICDAR.2015.7333910.

[16] Z. Zhao, H. Kang, B. Wang, C. He, Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception (2024). arXiv:2410.12628.

[17] J. Chen, R. Zhang, Y. Zhou, C. Chen, Towards aligned layout generation via diffusion model with aesthetic constraints, in: The Twelfth International Conference on Learning Representations, 2024, pp. 1–18.

[18] H. Cheng, P. Zhang, S. Wu, J. Zhang, Q. Zhu, Z. Xie, J. Li, K. Ding, L. Jin, M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis, in: 2023

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15138–15147. doi:10.1109/CVPR52729.2023.01453.

[19] J. Gu, X. Shi, J. Kuen, L. Qi, R. Zhang, A. Liu, A. Nenkova, T. Sun, Adopd: A large-scale document page decomposition dataset, in: The Twelfth International Conference on Learning Representations, 2024.

[20] M. Li, Y. Xu, L. Cui, S. Huang, F. Wei, Z. Li, M. Zhou, Docbank: A benchmark dataset for document layout analysis, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 949–960. doi:10.18653/v1/2020.coling-main.82.

[21] M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, Z. Li, Tablebank: Table benchmark for image-based table detection and recognition, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, 2020, pp. 1918–1925.

[22] H.-X. Lang, Y.-Y. Li, Y. Wang, H. Wang, J. Dong, An automatic topic-oriented structured text extraction method based on crf and deep learning, in: 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD), 2022, pp. 1408–1413. doi:10.1109/CSCWD54268.2022.9776155.

[23] H. Yang, W. Hsu, Transformer-based approach for document layout understanding, in: 2022 IEEE International Conference on Image Processing (ICIP), 2022, pp. 4043–4047. doi:10.1109/ICIP46576.2022.9897491.

[24] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, Layoutlm: Pre-training of text and layout for document image understanding, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1192–1200. doi:10.1145/3394486.3403172.

[25] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che, M. Zhang, L. Zhou, LayoutLMv2: Multi-modal pre-training for visually-rich document understanding, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 2579–2591. doi:10.18653/v1/2021.acl-long.201.

[26] D. M. Arroyo, J. Postels, F. Tombari, Variational transformer networks for layout generation, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13637–13647. doi:10.1109/CVPR46437.2021.01343.

[27] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence (2017) 1137–1149doi:10.1109/TPAMI.2016.2577031.

[28] A. Gu, K. Goel, C. Ré, Efficiently modeling long sequences with structured state spaces, in: The International Conference on Learning Representations (ICLR), 2022, pp. 1–31.

[29] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, X. Wang, Vision mamba: Efficient visual representation learning with bidirectional state space model, in: Forty-first International Conference on Machine Learning, 2024, pp. 1–11.

[30] P. Hu, Z. Zhang, J. Ma, S. Liu, J. Du, J. Zhang, Docmamba: Efficient document pre-training with state space model (2024). arXiv:2409.11887.

[31] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable detr: Deformable transformers for end-to-end object detection, arXiv preprint arXiv:2010.04159 (2020).

[32] X. Wang, Z. Zheng, J. Shao, Y. Duan, L.-J. Deng, Adaptive rectangular convolution for remote sensing pansharpening (2025). arXiv:2503.00467.
URL https://arxiv.org/abs/2503.00467

[33] Y. Pu, Y. Wang, Z. Xia, Y. Han, Y. Wang, W. Gan, Z. Wang, S. Song, G. Huang, Adaptive rotated convolution for rotated object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6589–6600.

[34] K. 2014Spatial, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (9) (2014) 1904–16.

[35] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 936–944. doi:10.1109/CVPR.2017.106.

[36] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision – ECCV 2014, 2014, pp. 740–755. doi:10.1007/978-3-319-10602-1_48.

[37] X. Zhu, S. Lyu, X. Wang, Q. Zhao, Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios, in: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021, pp. 2778–2788. doi:10.1109/ICCVW54120.2021.00312.

[38] C.-Y. Wang, A. Bochkovskiy, H.-Y. M. Liao, Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7464–7475. doi:10.1109/CVPR52729.2023.00721.

[39] G. Jocher, A. Chaurasia, J. Qiu, Ultralytics yolov8 (2023).

[40] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, G. Ding, Yolov10: Real-time end-to-end object detection (2024). arXiv:2405.14458.

A. Appendix

A.1. Annotations of PeKi Dataset

It is time-consuming and labor-intensive to manually annotate the images of various document types with complex layout categories. We employ about 12 annotators and 3 researchers to annotate these complex document images for about 4 months. The definition of categories and the guideline of annotations are carefully designed and can be basically applied to other types of documents. The layout annotations of the bounding boxes in our PeKi dataset are in standard MSCOCO [36] format for the classic detection task.

A.2. PeKi Dataset Label Definition

Formula: This category corresponds to 'Equation' in [5] and 'Formula' in [18, 11]. It primarily annotates mathematical formulas within document images. Unlike previous works, the annotation box must include the associated Formula-Caption, if present. This ensures that each formula is linked to its corresponding reference number, enhancing its traceability. If no reference number is available, only the formula itself is annotated.

Formula-Num: This category identifies the reference number or caption associated with a formula. If a formula has a corresponding reference number, it is annotated within the bounding box of the Formula category. The motivation for introducing this category is to enhance document parsing by preserving contextual relationships. Without this, formulas and their references might appear far apart in the document structure, leading to errors in parsing. Additionally, a single formula might be referenced

multiple times across different sections. By including Formula-Caption, we facilitate backtracking and accurate linking of references.

Figure: Similar to previous datasets [10, 11, 9], this category annotates image content within documents. If a figure contains subfigures, a single bounding box is used for the entire figure, rather than annotating each subfigure separately. Unlike existing approaches, the annotation box for Figure must also cover the Figure-Caption information.

Figure-Caption: This category marks the caption of a figure. The bounding box must be contained within the Figure annotation. By structuring the annotation in this way, we ensure that the description remains associated with the image, facilitating content matching and back-referencing in the text.

Reference: This category is used to annotate the contents section of a document.

Table: Similar to existing datasets [10, 11, 9], this category annotates tables within document images. However, unlike prior approaches, the annotation box must also include the Table-Caption.

Table-Caption: This category identifies table captions and must be contained within the Table annotation box. This ensures that table descriptions remain linked to their respective tables, facilitating accurate text replacement and back-referencing.

Footer: This category annotates footer information, including page numbers and footnotes at the bottom of a document page. Since footers often contain multiple page numbers, special attention is required in annotation.

Header: This category marks header information, including page numbers, headings, and other upper-margin content.

Doc-Title: 'Doc-Title' represents the document's title. In research papers, it corresponds to the main title. In books, it is usually the title of the book cover, while in other document types, it refers to prominently formatted text at the top of the document, often bolded or underlined.

Title-Id: This category annotates numerical identifiers associated with section titles, such as section numbers. Unlike previous works that focus on annotating entire titles, we emphasize indexing title numbers without considering the content itself. This allows for a more structured representation of hierarchical relationships within the document.

Title-NoId: This category is used for document headings that lack numerical identifiers, such as keywords, summaries, or standalone headings.

Title-Body: Previous works annotated titles under a single Title category, covering both numbering and textual content. However, this approach overlooked long-form titles that span multiple lines or contain visual emphasis (e.g., bold or colored text). To improve document layout analysis, we refine this approach:

Title-Id is used to annotate the section number.

Title-Body is introduced to capture the emphasized textual content of the title.

For multi-line titles, Title-Body is used to annotate all lines except the last, which is assigned to Title-Last.

Title-Last: This category annotates the last line of a multi-line title. If a title spans multiple lines, all preceding lines are labeled under Title-Body, while the final line is annotated under Title-Last.