Graphical Abstract

Mamba-YOLO: Adaptive Convolution for Document Layout Analysis

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

Highlights

Mamba-YOLO: Adaptive 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: Adaptive 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 YOLO framework with an adaptive rectangular convolution for the first time. This can effectively perceive different layout elements with 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.

Keywords: Document Layout Analysis, Mamaba, YOLO, Adaptive Rectangular Convolution

1. Introduction

As a core component of document understanding, the importance of Document Layout Analysis (DLA) has become increasingly prominent[1, 2, 3]. Over the past few years, DLA has seen significant advancements. However, the existing work[4, 5] 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[5] adopt a multimodal approach, introduce a self-supervised training mechanism, and sense document layout based on Transformer[6] 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[7].

However, the current research faces three challenges. Firstly, there is a notable absence of comprehensive Chinese datasets that span multiple domains. Existing studies predominantly rely on English-language datasets, and even when Chinese data are available, they are often limited to specific fields or contain only a small proportion of Chinese content. 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 standard reports, government announcements, internal organizational notifications, and academic papers, which we have named the PeKi dataset. Compared to previously available datasets[8, 9, 10, 11], PeKi annotate and refining all

the "title" and focused on the fine-grained information. At the same time, the picture, table and equation areas cover the caption. For object detection, YOLO[12] is an outstanding algorithm known for its outstanding accuracy and speed on natural images. To enhance YOLO's performance on DLA, we adopt the efficient design principles of VMamba[13] and integrates 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 Adaptive Rectangular Convolution(ARConv), especially for objects with varying aspect ratios. For feature loss in continuity detection, especially when dealing with sequential titles, we introduce Mamba into the network to build a linear relationship to extract continuity.

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 Adaptive Rectangular Convolution.
3. Our model effectively mitigates domain transfer issues, significantly improving baseline performance on benchmark suites. On several existing open-source datasets CDLA[8], PubLayNet[9], DocLayNet[10], and D4LA[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[9] 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[8] dataset is a Chinese document layout analysis dataset tailored for scholarly publication, a single domain. DocLayNet[10] manually annotates four languages and six document types, comprising 11 categories of layout analysis from diverse document types. D4LA[11] 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, 5, 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,

Dataset	Image	Class	Instance	A.M.	Source Format	Document Type	Language
PubLayNet[9]	360,000	5	3,311,660	Automatic	PDF	Articles	English
CDLA[8]	6,000	10	70,928	Manual	PDF	Articles	Chinese
DocLayNet[10]	80,863	11	1,107,470	Manual	PDF	Financial Reports,Manuals,Scientific Articles,Laws,Regulations,Patents,Government Tenders	English,German,French,Japanese
D4LA[11]	11,092	27	146,846	Automatic	PDF	Ideological, Moral Cultivation,Basic Law Education	English
M6Doc[18]	9,080	74	237,116	Manual	PDF,Scanned,Photographed	Scientific Articles,Textbooks,Books,Test papers,Magazines,Newspapers,Notes	English,Chinese
DocBank[20]	500,000	13	-	Automatic	PDF	Articles	English
TableBank[21]	4,000	1	417,234	Automatic	Latex, Word	Articles	English
DocSynth-300K[16]	11,314	10	109,004	Automatic	PDF	Academic,Textbook,Market Analysis,Financial	Chinese
PeKi(Ours)	46,777	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. A.M. denotes the annotating means.

Category	equation	equationcaption	figure	figurecaption	reference	table	tablecaption
Training/Validate/Test	285/19/46	273/20/45	7682/921/928	3926/515/496	730/78/93	16798/2052/2127	7375/929/930
Category	titlecover	footer	header	titlewithoutindex	titleindex	titlenext	titlecontent
Training/Validate/Test	3451/420/392	26498/3357/3307	13786/1728/1725	1648/245/202	214037/27951/26565	1060/151/117	587/79/82

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

Global-to-Local Design-Controllable Receptive Module enables models to efficiently detect targets at different scales. DiT[5] 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

From Tab. 1, current datasets are English[9, 11], Chinese[8], or multi-lingual[10, 18]. Above data sets lack not only fine-grained layout analysis, but also Chinese multi-

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

domain data sets. In response to this, we introduce the PeKi dataset, as shown Fig. 1. PeKi dataset is collected from multiple channels, including journals in computer science(J), university textbooks(B), industry standard reports(S), government reports(R), and organizational notifications(N). Number of instances corresponding to each annotation type is presented in Tab. 2.

From Fig. 1 and Tab. 2, we focus on the fine-grained information of "title"[11, 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: "titleindex" indicates the serial number of the title and does not include any subsequent text information; "titlewithoutindex" is used for titles that are in bold but do not have an index; "titlecontent" is for titles that are both indexed and in bold; "titlecover" represents the title information on the document cover; "titlenext"

is used when the "titlecontent" is too long and the title information is wrapped to the next line. Like CDLA [8], we also pay attention to the fine-grained information of table caption, figure caption and equation caption, captions annotation area needs to be within the scope of the picture, table or equation annotation. These refinements enable a more nuanced and accurate representation of title information, enhancing the overall quality and usability of the dataset.

The PeKi dataset distinguishes itself from CDLA [8], PubLayNet [9], DocLayNet [10] and D4LA [11] in the following aspects:

1. Refine all the title in the document.
2. Place the annotation ranges for figure captions, table captions, and equation numbers within their respective object boundaries.
3. Discard the "Text" category used in previous work.
4. Within each sentence led by a title, only the title index is annotated, without considering the corresponding textual information.

4. Method

4.1. Model Architecture

In order to overcome varying aspect ratios and loss of continuity target features, we propose a novel method named DocMY. Overall structure of DocMY is illustrated in Fig. 2, where a Generalized ELAN State Space Model Block (GESSM Block) is integrated into the YOLOv9[12] framework with ARConv. ARConv kernels are employed for feature extraction, thereby better accommodating objects with inconsistent aspect ratios in documents, see 4.2 for more details. Inspired by the GELAN design in YOLOv9, we optimizes the SSM network structure design by introducing the GESSM Block module. Within the existing efficient layer aggregation network, this leverages the inference capabilities of Mamba, making the model more efficient and lightweight, see 4.3 for more details.

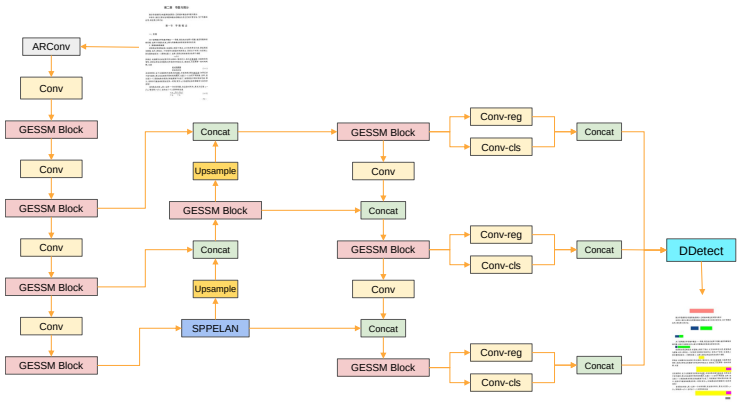

Figure 2: DocMY Model Architure

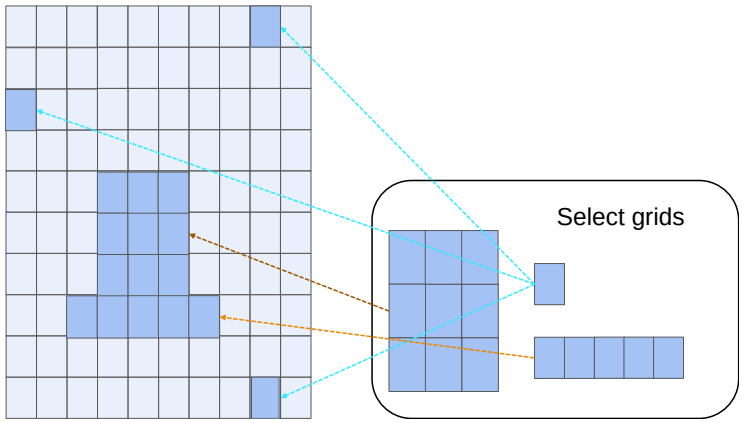

Figure 3: Adaptive Rectangular Convolution

4.2. Adaptive Rectangular Convolution

Convolutional networks primarily use convolutional kernels for feature extraction from images, iterating over the original input image using (k, k) kernels. In this paper, Adaptive Rectangular Convolution kernels(m,n;m<=n)[31], as shown in Fig. 3, are utilized through convolution operations, making them better suited for document layout analysis and capable of more accurately capturing the layout information of document. The design of convolution kernels with inconsistent aspect ratios reduces the likelihood of text boxes being split apart.

Research has found that setting select grids (1,1), (3,3), (1,5) convolution kernel can effectively enhance the performance of DLA. For elements within documents such as figure and table, where the aspect ratio is close to 1:1, the original (3,3) convolution operation is used. For figure caption, table caption, header, and titlecover, which have the largest aspect ratios, a (1,5) convolution kernel operation yields better results. For titleindex, we use (1,1) for feature extraction. Under the influence of heuristic rules, rules are designed to make ARConv better adapt to different positions and perceive the feature information of regions in the document according to different structure layouts in the document.

4.3. Generalized ELAN State Space Model

To mitigate the feature loss problem and parameters problem, inspired by Vision Mamba and VMamba, we applies Mamba to the visual tasks of DLA and integrates it with YOLOv9[12] which has a special module Generalized ELAN, as shown in Fig. 5. To address the feature loss of continuity information, traditional approaches concatenate the inputs from top to bottom and left to right. However, in this current situation, there are continuous vertical blocks. If the previous method were still used, it would not efficiently extract features and loss features. Therefore, we alter the current concatenation method by using a left-to-right, top-to-bottom concatenation approach. The specific operational process is illustrated in Fig. 4. This method is merged with the concatenation techniques of prior work, effectively extracting feature information.

Retaining the capabilities of Generalized ELAN, and drawing inspiration from the VMamba module design, we introduce the State Space Model module into GELAN. The inclusion of the residual structure facilitates the perception of special areas in document layout analysis and helps avoid the loss of features caused by down sampling. Fig. 5 shows the specific structure of GESSM Block proposed in this paper.

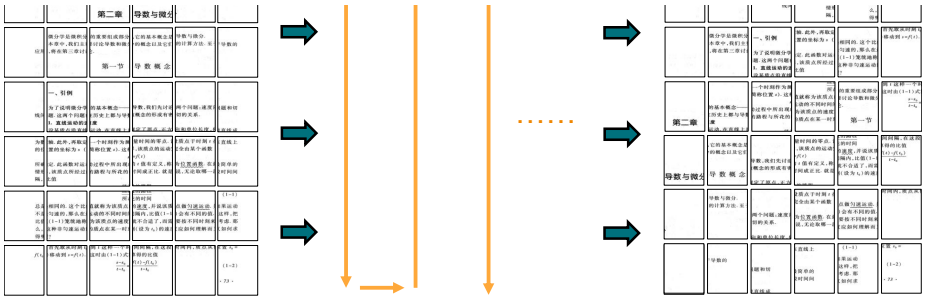

Figure 4: Scanning Mechanism

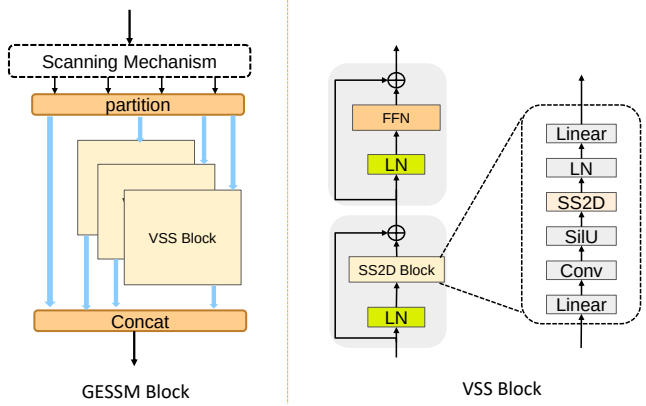

Figure 5: Generalized ELAN State Space Model

5. Experiments

5.1. Experimental metrics and datasets

For evaluation metrics, we report COCO-style[32] Precision, Recall, mAP-50, mAP50-95, and parameters. Experiments in this paper are based on the PeKi dataset to validate the effectiveness and superiority of the DocMY. Additionally, experiments are conducted on currently popular and publicly available document layout analysis benchmarks, including CDLA[8], PubLayNet[9], DocLayNet[10], and D4LA[11].

5.2. Experimental Details

Hardware configuration for experiments is as follows: Intel(R) Xeon(R) Silver 4216 CPU @ 2.10GHz; system memory of 128GB; and NVIDIA GeForce RTX 2080

	Method	P(%)	R(%)	mAP50(%)	mAP50-95(%)	parameters
Multimodal	Layoutlmv3[4]	92.8	73.9	94.2	-	133,267,834
	DiT[5]	89.7	86.4	91.4	84.3	87,679,409
Unimodal	YOLOv5[33]	91.6	**94.3**	95.6	77.2	21,775,401
	YOLOv7[34]	92.2	93.1	96.0	81.3	22,194,944
	YOLOv8[35]	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[36]	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.

Ti with 11GB of VRAM. All models mentioned in this paper were trained using a strategy that starts from scratch, with a total of 100 training epochs. Initial learning rate is set to 1e-2, and the final learning rate is set to 1e-5. For setting the learning rate, we use linear warm-up during the first three epochs, and for subsequent epochs, the corresponding decay method is set according to the model scale. In the last 15 epochs, mosaic data augmentation is turned off.

5.3. Results on PeKi Datasets

As a Chinese dataset, PeKi fills the gap in current document layout analysis efforts focused on the Chinese language. Experiments are conducted comparing Unimodal YOLO and Multimodal Transformer-based architectures such as LayoutLMv3[4] and DiT[5], as well as VMamba[13] and Vision Mamba[29]. Experimental results of different methods on our dataset are presented in Tab. 3.

It can be observed that the model design presented in this paper achieves significant performance improvements compared to current method. When compared with the YOLO series of works, under approximately the same number of parameters, our model reaches state-of-the-art performance on PeKi. Compared to the baseline[12], our model improves Precision by 3.2% and Recall by 4.1%. By introducing the State Space Model method, the issue of continuous regions of interest in documents is effectively alleviated, allowing for the capture of title targets.

Datasets	Method	P(%)	R(%)	mAP50(%)	mAP50-95(%)
PubLayNet[9]	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[8]	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[10]	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[11]	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 4: Effects of Our Method on PubLayNet, CDLA, DocLayNet, and D4LA. **Bold** indicates performance improvement compared to baseline.

Method	Header	Text	Reference	Figure caption	Figure	Table caption	Table	Title	Footer	Equation	mAP
YOLOv9	92.4	96.9	96.4	92.3	97.0	95.1	98.9	95.1	83.9	92.0	94.0
Ours	93.8	98.0	97.3	94.7	97.7	98.1	99.4	97.2	88.2	96.7	96.1

Table 5: Performance comparisons on CDLA dataset.

5.4. Results on Other Datasets

Experiments are also carried out on other datasets to verify the validity of the proposed DocMY. Tab. 4 shows the experimental results of this paper's network structure in PubLayNet[9], CDLA[8], DocLayNet[10], and D4LA[11]. As shown in tab. 4, proposed model exhibits varying degrees of improvement over the baseline. On the PubLayNet[9] dataset, although the SOTA is 96.2%, but our model achieves very competitive results with fewer parameters. For the DocLayNet[10] document dataset, the accuracy in recognition improves by 1.0%, and the mAP50-95 performance increases by 1.1%.

5.5. Results on CDLA

Tab. 5 demonstrates that DocMY achieves State-Of-The-Art performance compared to the YOLOv9 model across all categories in the Chinese CDLA dataset[8]. This differential performance can be attributed to GESSM Block, which effectively captures content such as title and text that are easily lost in previous work. Additionally, the design of ARConv, which capture regions of interest with varying aspect ratios.

Method	P(%)	R(%)	mAP50(%)	mAP50-95(%)
Baseline[12]	91.6	89.9	94.1	81.9
B+GESSM	93.5	93.9	96.9	81.9
B+GESSM+ARConv	94.8	94.0	96.5	83.3

Table 6: Ablation study of model Architure.

5.6. Ablation Experiment

We conduct thorough experiments to validate the effectiveness of GESSM Block and ARConv. Tab. 6 summarizes the ablation experiments conducted on our proposed PeKi dataset.

From Tab. 6, it can be observed that the baseline model performs with a precision of 91.6%, recall of 89.9%, mAP50 of 94.1%, and mAP50-95 of 81.9%. These results indicate that the baseline model already possesses strong object detection capabilities. However, addition of ARConv and GESSM alleviates the problem of inconsistent aspect ratio and feature loss, model performance is stronger.

Effect of introducing the State Space Model enhancement strategy (B + GESSM): Integrating the State Space Model enhancement strategy into the baseline model boosts the Precision to 93.5%(+1.9%) and the recall to a remarkable 93.9%(+4.0%). This indicates that the State Space Model strategy effectively enhances the models ability to identify and locate targets, primarily due to its scanning mechanism and linear perceptual capabilities.

Effect of further integrating ARConv (B+ GESSM + ARConv): Results show an increase in Precision to 94.8%(+1.3%) and recall to 94.0%(+0.1%), suggesting that ARConv may refine feature representations, achieving more precise object detection while maintaining model stability. Specifically, the mAP50-95 climbs to 83.3%, a significant improvement over the previous two models.

In summary, this ablation study demonstrates that incrementally introducing the State Space Model and the ARConv module effectively improves the overall performance of DLA, providing valuable insights for further optimization and the construc-

tion 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 figures, tables with captions, equations, 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 (SSM) and Adaptive Rectangular Convolution (ARConv). 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 pictures, tables, and formulas with content in the body. DocMY leverages the SSM and ARConv to address limitations in traditional approaches to complex document layouts. However, there still exist some weaknesses of this work. First, the PeKi dataset primarily covers Chinese documents. Second, the current work is not able to effectively analyze targets that are too large. In the future, we plan to expand the PeKi dataset by including a wider variety of document types and languages. Also, it is interesting to employ multi-modality to enrich the input features for document layout analysis.

Acknowledgments

This work is partly supported by National key r&d program (Grant no. 2019YFF0301800), 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] 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.

[6] 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.

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

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

[9] 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.

[10] 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.

[11] 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.

[12] C.-Y. Wang, I.-H. Yeh, H.-Y. Mark Liao, Yolov9: Learning what you want toălearn 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] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 764–773. doi:10.1109/ICCV.2017.89.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.