2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT) | 979-8-3503-9905-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICEICT57916.2023.10245716

Document image object detection algorithm based on Transformer and Mixed-MLP Network*

*Note: Sub-titles are not captured in Xplore and should not be used

1st Honghong Zhang
dept. School of Information Science and Technology
Qingdao University of Science and Technology
Qingdao, China
may_i zhh@163.com

2nd Canhui Xu
dept. School of Information Science and Technology
Qingdao University of Science and Technology
Qingdao, China
ccxu09@yeah.com

3rd Yuanzhi Cheng
dept. School of Information Science and Technology
Qingdao University of Science and Technology
Qingdao, China
yzcheng2007@163.com

***Abstract*—The feature extraction network based on convolutional neural network has a strong ability to extract local features, but it is insufficient to connect the information of different spatial regions in global scope. To solve this problem, a document image object detection model using Transformer as the feature extraction backbone network is proposed. On the one hand, Transformer extracts image features and utilizes self-attention mechanism to establish global information connections. On the other hand, Mixed-MLP network consisting of channel mixing (C-Mix) layers and spatial mixing (S-Mix) layers explores information communication in spatial dimension and channel dimension. The extracted image features are represented by multi-scale hierarchical network, and feature pyramid is built to detect objects of different scales. The experimental results show that the precision of detection on PubLayNet and ICDAR-POD are 0.958 mAP and 0.952 mAP respectively, which achieve competitive results.**

***Index Terms*—deep learning,object detection,document image,Transformer,layout analysis**

I. Introduction

Document image object detection is a process of geometric segmentation and logical understanding of page elements such as text, formula, figure, table, and list, etc. It segments page into different regions, and then realize logical recognition according to the content of the regions. It is also the pre-step of document retrieval, text recognition, text segmentation and other automatic applications.

Traditional document object detection methods mainly fall into two groups: bottom-up and top-down methods. The former [1] is gradually connected from a small image region to the larger, and the calculation is relatively complex. The latter [2] divides the image from large to small until it reaches the scale of page elements. Generally speaking, traditional methods rely on complex handmade features and heuristic rules, which is difficult to guarantee generalization for document images with complex layouts and diverse page elements.

Compared with traditional methods, deep learning model has a strong feature extraction and representation ability. Convolutional neural network (CNN) is able to extract more detailed features with complex structure and deep layers. Many document image object detection methods adopt CNN as backbone network to extract image features. The method proposed by He et al. [3] uses VGG Net [4] to extract image feature information, and Fully Convolutional Network (FCN) [5]for page segmentation and table detection. In [6], they compared the influence of network size on the performance of detection models based on AlexNet [7] and VGG-16 [4]. As a milestone innovation of CNN, ResNet [8] achieves better performance of layout analysis algorithms by establishing multi-scale feature output in [9] [10].

CNN focuses on the correlation of local information and has a strong ability to extract local features. For object detection task, the global context information of image is significant. Based on the success of Transformer in NLP, some efforts have attempted to introduce Transformer into computer vision tasks. The self-attention mechanism [12] is used to improve the ability to connect global context information of network. In [13] [14], they focus on the interdependence of pixels in global field and obtain global context information by convolutional calculation and matrix multiplication. There are some methods [15] [16] replaced multiple the convolutional layers with self-attention layers in ResNet network, which evaluated the effectiveness of self-attention mechanism in

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61806107 and 61702135, Shandong Key Laboratory of Wisdom Mine Information Technology, and the Opening Project of State Key Laboratory of Digital Publishing Technology.Corresponding author: Canhui Xu. (email: ccxu09@yeah.com)

image feature extraction. ViT [17] is the first attempt to apply pure Transformer framework to visual classification tasks. It split an image into fixed-size patches as similar to word vectors of sentences in language tasks. ViT performs well in multiple image processing tasks, but it focuses more on the connection of local information. In addition, large-scale training data is required for good performance. Therefore, DeiT [18] adopts distillation way which made models based on Transformer learn inductive bias of CNN models, and improves the processing ability of image data. DETR [19] chooses to combine CNN with Transformer. The input of Transformer network is the image features extracted by CNN. In order to obtain global context information, Liu et al. [20] limit self-attention calculation into non-overlapping local windows with fixed size, and realize global information interaction by constantly shifting window locations for cross-window connection.

Hence, there are some researches focus on Transformer of Natural language processing (NLP), which utilize self-attention mechanism to explore global feature dependency of images. LayoutLM [11] is one of the very first attempts to design a pre-training model for document image understanding based on Transformer layers. LayoutLM 2.0 [30] is a multimodal Transformer encoder network with spatial awareness self-attention mechanism, which accepts text, image, and layout as the input and improves the performance of document image understanding. In this paper, our proposed object detection model pay attention to explore connection among different regions image features in global scope. The image feature extraction network is based on Transformer to improve the perception ability of global feature information. In addition, Multi-Layer Perceptron (MLP) structure is used to design a Mixed-MLP network, which integrates information in spatial and channel dimensions. The works of our paper are summarized as follows:

- The backbone network is established based on Transformer. Each stage consists of multiple Transformer blocks and each Transformer block realizes the computation of self-attention mechanism. The output features build a multi-scale feature pyramid which adapts to the detection of page elements with different scales.
- Mixed-MLP network is embedded between backbone network and multi-scale feature pyramid. With feature map patches as input, Mixed-MLP network allow communication in channel dimension with channel-mixing layers, and realize communication in spatial dimension with spatial-mixing layers. Mixed-MLP focus on emphasizing information communication in spatial and channel dimension.

II. Method

The overall framework of document image object detection based on Transformer is shown in Fig. 1. Swin Transformer is introduced as an image feature extraction network, which is a four-stage structure with multi-scale outputs. Each stage is composed of several pairs of Swin Transformer blocks. The output feature maps are split into image patches by Tokensize operation. Mixer-MLP network takes as input a sequence consisting of multiple non-overlapping image patches. Mixer-MLP consists of multiple layers of identical size, and each layer consists of two MLP blocks. It aims at allowing communication between different channels and spatial locations. The multi-scale feature pyramid is constructed by feature maps outputted by Mixer-MLP network. The detailed information of each part is described as following sections.

A. Transformer feature extraction network

In order to obtain information communication within channel and spatial dimension of images, Swin Transformer network is adopted as backbone network to extract image features and build multi-scale feature representation pyramid. Compared with other visual Transformer structures, Swin Transformer takes image patch partition as its input. The self-attention is computed within each non-overlapping window and global information connections is realized by shifting windows. As shown in Fig.2, a Swin Transformer block includes a Multi-head Self-Attention (MSA), a Multi-Layer Perceptron (MLP). A LayerNorm (LN) layer [21] is applied before each MSA module and each MLP, and a residual connection is applied after each module. Each MLP module includes two fully connected layers and GELU [22] activation function. A skip connection is also applied after each MSA and MLP layer. The computation of global self-attention mechanism is unaffordable for an image with large resolution. Hence, Swin Transformer designs a shifted window mechanism to reduce the complexity of computation.

As illustrated Fig.2, the global self-attention computation is realized in two successive Swin transformer blocks. In the first block, window based multi-head self-attention(W-MAS) means the self-attention score is only computed among patches in the same window. In this case, the self-attention mechanism is limited in each local window and lacks connections across windows. The next block including shifted window based multi-head self-Attention(SW-MSA) layer realizes windows shifting based on last block. The patches belonging to same window in last block are displaced into different Windows. Then, the self-attention mechanism is computed again to among the patches within shifted window. The shifting window strategy extends the self-attention computation within local windows to the global scope of the image. With the shifted window partitioning approach, consecutive Swin Transformer blocks are computed as:

$$\begin{aligned} \widetilde{p}^{l} &= \mathrm{W-MSA}\left(\mathrm{LN}\left(p^{l-1}\right)\right) + p^{l-1}, \\ p^{l} &= \mathrm{MLP}\left(\mathrm{LN}\left(\widetilde{p}^{l}\right)\right) + \widetilde{p}^{l}, \\ \widetilde{p}^{l+1} &= \mathrm{SW}-MSA\left(\mathrm{LN}\left(p^{l}\right)\right) + p^{l}, \\ p^{l+1} &= \mathrm{MLP}\left(\mathrm{LN}\left(\widetilde{p}^{l+1}\right)\right) + \widetilde{p}^{l+1}. \end{aligned} \tag{1}$$

Where $\widetilde{p}^{l}$ represents the features of W-MSA or SW-MSA output in the l_{th} block, and p^{l} represents the output features of MLP layer in the l_{th} block. The design of shifted window mechanism requires that Swin Transformer blocks always connect in pairs. The number sequence of Stage1 to Stage4

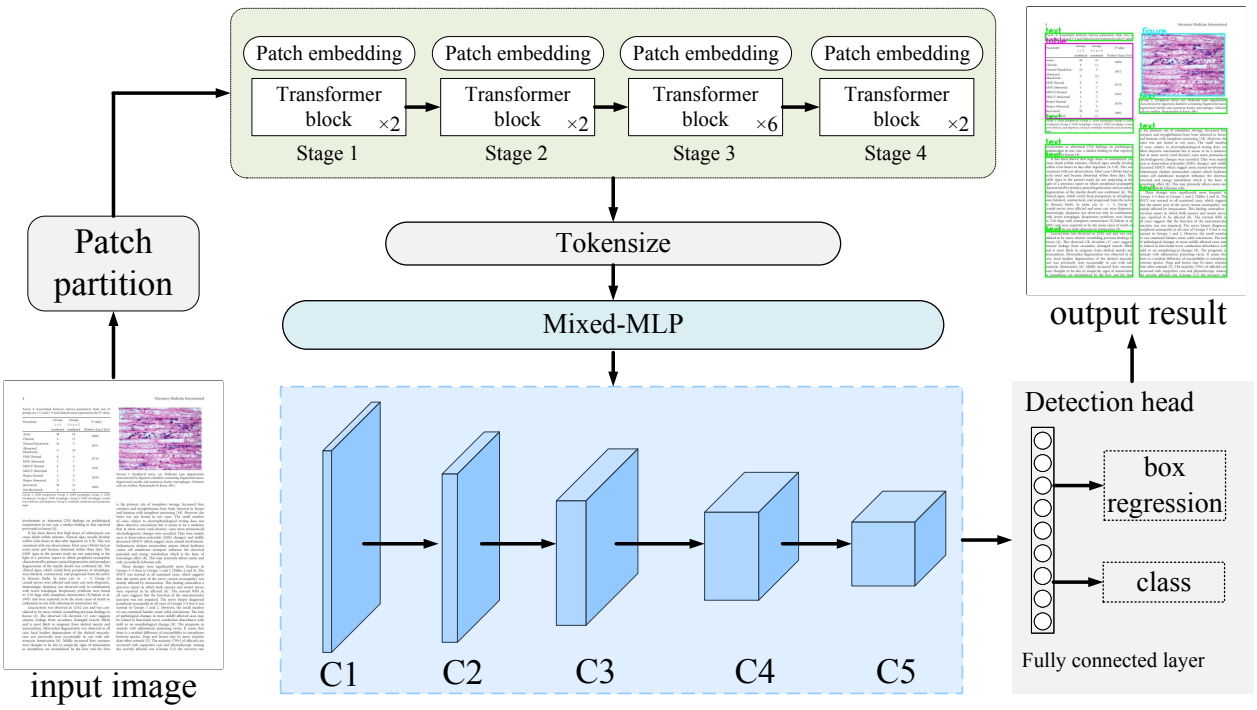

Fig. 1. The overall architecture of our network.

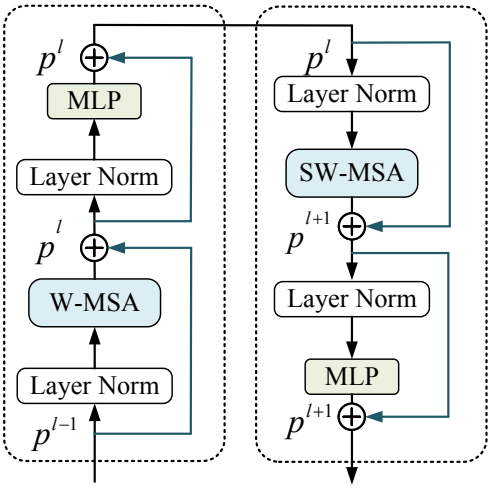

Fig. 2. Two successive Swin Transformer blocks.

shown in Fig. 1 is (2,2,6,2), which is the miniature version of Swin Transformer (SwinT). The formula of self-attention computation is as follows:

$$G_{attntion}\left(Q,K,V\right)=\text{softmax}\left(QK^{T}\Big/\sqrt{d}+B\right)V \quad (2)$$

Where $Q\in R^{M^2\times d}$,$K\in R^{M^2\times d}$ and $V\in R^{M^2\times d}$ represent query vetor, key vector and value vector in Transformer respectively, d equals $query/key$. M^2 is number of patches in the window, and B is a relative position bias. The self-attention mechanism is used to build the relationship between image patches, and extract image features containing more context information.

B. Mixed-MLP Network

Self-attention mechanism convolution operation are able to establish information connections in a global or local scope. In this paper, MLP is used to design Mixed-MLP networks for information communication in channels and spatial dimensions respectively.

As shown in Fig.3, Mixed-MLP network includes channel-mixing(C-Mix) and Spatial-mixing(S-Mix) layers which are implement by MLP layers.The output feature map of backbone network with resolution of (H,W) is the input of Tokensize module. The feature map is split into non-overlapping patches of same size in spatial dimension. The size of each patch is $p\times p$, and there are Npatches,where $N=HW/p^2$. Then, all patches are projected linearly with same projection matrix and input as a sequence $S\in R^{C\times N}$ into the following Mixed-MLP network. Each row of S is a linear projection of a patch. The formula is as follows:

$$L=\{l|Tokensize\left(p\right)\} \quad (3)$$

$$S=\text{Proj}\left(L\right) \quad (4)$$

Where L represents the set of patches after $Tokenize$ operation. Mixed-MLP network allow communication between different spatial locations and different channels. Firstly, each row of S is inputted to a single C-Mix layer, which obtain

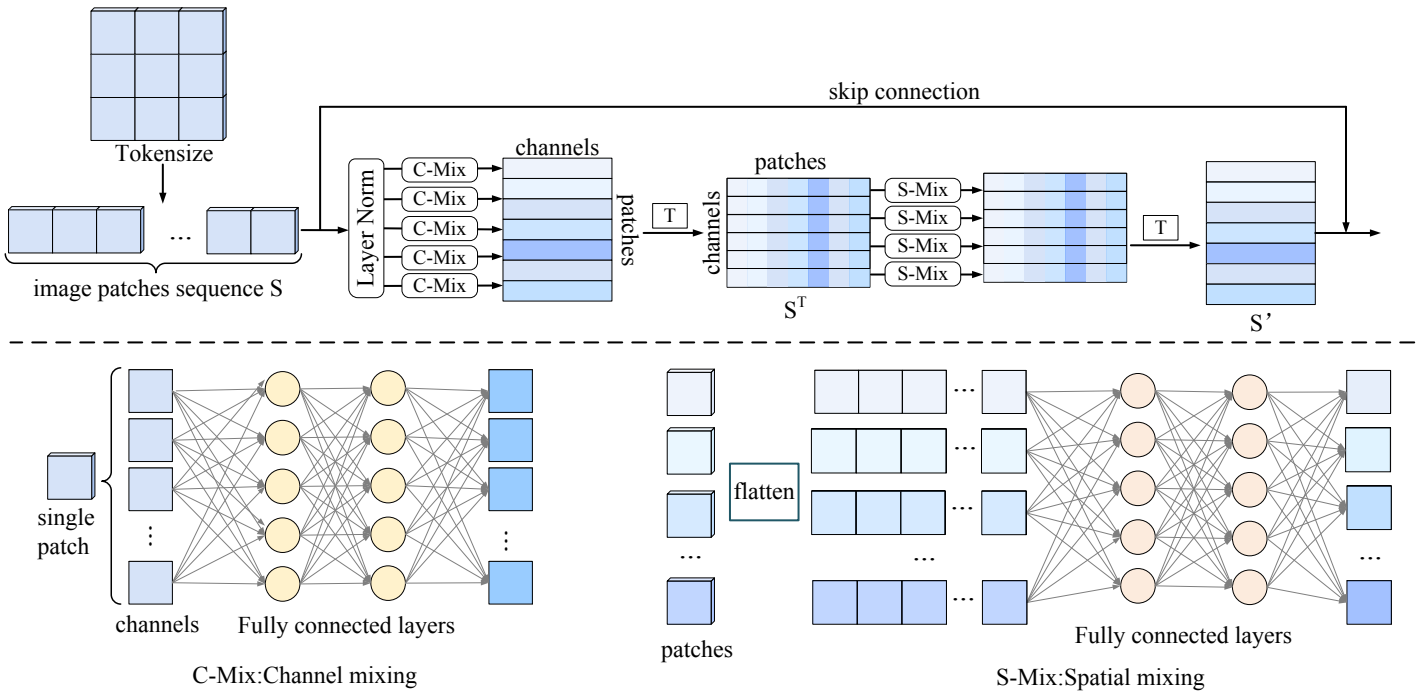

Fig. 3. The structure of Mixed-MLP network.

the information communication within a patch. Then, S is transposed and it can be seen that the input of S-Mix layer is the row of transposed S. Therefore, the information of patch channels can be connected. Finally, the S^T is transposed to obtain a feature map S' which has same resolution with original feature map S. The formula can be written as follows:

$$S^T = \left[w_2 \varphi w_1 \left(\mathrm{LN}\left(S\right)\right)\right]^T \tag{5}$$

$$S' = \left[w_4 \varphi w_3 \left(\mathrm{LN}\left(S^T\right)\right)\right]^T \tag{6}$$

where LN represents LayerNorm [21] layer and φ represents GELU [22] nonlinear activation layer. In addition, there is a standard architectural component: skip-connections from initial input features S to the output features Y:

$$Y = S \oplus S' \tag{7}$$

Y is the feature output through Mixed-MLP network.

Document images have diverse page elements with different scales. Hence, a multi-scale feature pyramid is established to adapt to the detection of different scale objects. Mixed-MLP network is embedded in the connection between backbone stage and layer of the feature pyramid. The formula is described as follows:

$$C_i = C_{i+1} \oplus G_{Mixed-MLP}\left(Z_i\right), i = 3, 2, 1 \tag{8}$$

Which Z_i represents the i_{th} output feature map of the backbone network stage. C_5 is obtained by down-sampling C_4. In this way, a multi-scale feature pyramid containing five layers of feature maps is obtained, which is used to detect document objects across scales.

III. Experiments

A. Datasets

In this section, experiments are conducted on available publicly datasets PubLayNet [10] and ICDAR-POD [23] to evaluate the effectiveness of our network.

- PubLayNet is a large document dataset with more than 300,000 document images. Its content is based on over one million PDF articles publicly available on PubMed Central™. The data set annotates five types of document layout elements with bounding boxes and polygons: text, title, list, illustration, and list. The evaluation metric is as same as COCO competition which is the mean average precision (AP) @ Intersection over Union (IoU) [0.50:0.95].
- The ICDAR-POD dataset contains 2,412 document images scanned from contemporary magazine and technical articles. The data set is annotated with three classes: formulas, tables, and illustrations.

B. Implement details

The experiment environment of is Ubuntu 20.04 system. All the models are implemented under the PyTorch framework and trained on one Nvidia GTX 2080Ti GPU by the Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and weight decay of 0.0001. PubLayNet is trained 3 epochs at a basic learning rate of 0.0009, which is divided by 10 every epoch. The training epochs of ICDAR-POD are 30, which is divided by 10 every 10 epochs with the basic learning rate of 0.0025. All of the experiments are conducted on an

Ubuntu 20.04 workstation with an Intel(R) Xeon(R) Silver 4210 2.20GHz CPU 64GB RAM.

C. Results discussion

TABLE I
PERFORMANCE COMPARISONS ON PUBLAYNET DATASET.

Method	Text	Title	List	Table	Figure	mAP
Mask R-CNN [24]	91.6	84.0	88.6	96.0	94.9	91.0
ATSS [25]	**96.9**	84.3	94.3	97.1	94.0	93.3
CDDOD [26]	92.3	91.3	91.3	93.4	92.7	92.2
YOLOACT++ [27]	90.9	62.1	85.1	88.0	91.4	83.5
VSR [28]	96.7	**93.1**	94.7	97.4	96.4	95.7
Our Method	96.8	91.4	**95.8**	**98.1**	**97.0**	**95.8**

1) PublayNet dataset: PubLayNet is one of the largest document image data sets, containing nearly 360,000 document page images. As shown in TableI, our method has the best performance in the categories of list, table and figure, reaching 95.8%, 98.1% and 97.0% AP respectively. Among them, the list category has higher increase with 1.1%AP. A list object includes both the text content and the small bullets or numbers, such as black dot symbol, numeric number or letter number. This kind objects needs to fully consider the global information in spatial dimension. When extracting image features, Swin Transformer network focus on capture global context information and the shifting window mechanism allows the information communication among different regions of image. In addition, Mixed-MLP network is capable of establishing the relation in spatial dimension and channel dimension information through computation of MLP layer. They are contribution to the information interaction in a global field.

2) ICDAR-POD dataset: In general, our method achieves high precision on the ICDAR-POD dataset. When the threshold of IoU (Intersection over Union) was 0.6 and 0.8, mAP reaches 95.2% and 91.2%, respectively. When the IoU equals 0.6, the results increases by 10.7% AP on Formula, 2.3% AP on Table, 1.9% AP on Figure. When the threshold of IoU increases to 0.8, the categories of formulas and tables increase by 7% and 2.5%AP, respectively. It can be observed that our model achieves better performance on Formula category. ICDAR-POD dataset detects three objects: formulas, tables, and illustrations. Among them, formula has obvious appearance compared with other text regions.As can be seen from TableII, in an internal comparison of the same method, precision of formula is lower than other two categories. One of the factors is that the formula object consists of two parts: formula and formula number, which may result in some inaccurate detection box. Mixed-MLP network can not only consider the relation within image patches in spatial dimension, but also emphasize the feature relation within the formula to detect the object. It is necessary to capture the context relation between formula and formula number. Mixed-MLP provides information communication and the features of different detection objects can be distinguished. The detection precision of each category can be improved.

3) Ablation study: In this section, ablation experiments conducted on PubLayNet are designed to evaluate the effect of different number of Mixed-MLP networks to the detection precision. As shown in TableIII, the number of Mixed-MLP increase from 0 to 8 with a step of 2. Intuitively, when the number increases to 8, the method achieves highest performance. When the number is between 0 and 6, the detection accuracy is slightly improved compared with the benchmark model. When the number of embedded Mixed-MLP is small, it is not sufficient to establish internal and external information communication of images relying on fewer MLP modules. Convolution operation, due to the limitation of convolution kernel size, has a strong extraction ability for image local features. While previous self-attention mechanism contribute to capture long dependency of image information, it computational complexity is quadratic for the image with higher resolution. Compared with convolution operation and self-attention, MLP mainly relies on the fully connected layers within the feature map patches. A few MLP layers is inadequate to build contextual connections in global scope. After MLP layer computation, the information of each image patches of the image can be fully integrated, and explore the relationship in different channels and spatial regions.

IV. CONCLUSION

This paper proposed a document object detection model which utilizes Transformer as feature extraction network and integrates Mixed-MLP network. Transformer is able to explore feature representation in global scope, and Mixed-MLP enhances information communication in spatial and channel dimension. The self-attention mechanism and MLP operation explore the connection of information in non-local field of image and aggregates global contextual information. The experiment results demonstrate that our method improves the precision of document object detection, which is expected to provide inspiration for following works on document image object detection based on Transformer.

REFERENCES

[1] L. O'Gorman, The document spectrum for page layout analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1993, 15(11): 1162-1173

[2] G. Nagy, S. Seth, M. Viswanathan, A prototype document image analysis system for technical journals[J]. Computer, 1992, 25(7): 10-22.

[3] He D, Cohen S, Price B, et al. Multi-Scale Multi-Task FCN for Semantic Page Segmentation and Table Detection[C]//2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Washington DC: IEEE Computer Society, 2017: 254-261.

[4] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[C]//International Conference on Learning Representations (ICLR). Piscataway, NJ: IEEE, 2015: 1-14.

[5] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(4): 640-651.

[6] Yi X, Gao L, Yuan L, et al. CNN Based Page Object Detection in Document Images[C]//2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Washington DC: IEEE Computer Society, 2017: 230-235.

[7] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

TABLE II
PERFORMANCE COMPARISONS ON ICDAR-POD DATASET.

	AP(IoU=0.6)				AP(IoU=0.8)			
Method	Formula	Table	Figure	mAp	Formula	Table	Figure	mAP
NLPR-PAL	83.9	93.3	84.9	87.4	81.6	91.1	80.5	84.4
icstpku	84.9	75.3	67.9	76.0	81.5	69.7	59.7	70.3
FastDetectors	47.4	92.5	39.2	59.7	42.7	88.4	36.5	55.9
Vislnt	52.4	91.4	78.1	74.0	11.7	79.5	56.5	49.2
sos	53.7	93.1	78.5	75.1	10.9	73.7	51.8	45.5
UITVN	19.3	92.4	78.6	63.4	6.1	69.5	55.4	43.7
Matiai-ee	11.6	78.1	32.5	40.7	0.5	62.6	13.4	25.5
HustVision	85.4	93.8	85.3	88.2	29.3	79.6	65.6	58.2
li eta [29]	87.8	94.6	89.6	90.7	86.3	92.3	85.4	88.0
Our Method	**97.1**	**96.9**	**91.5**	**95.2**	**93.3**	**94.8**	**85.6**	**91.2**

TABLE III
ABLATION EXPERIMENTS ON PUBLAYNET DATASET.

Method	Text	Title	List	Table	Figure	mAP
Baseline	92.7	86.9	92.6	97.2	95.8	92.6
×2	92.9	88.8	93.5	97.3	96.3	93.2
×4	93.0	88.5	93.5	97.3	96.2	93.3
×6	93.3	88.8	94.2	97.8	96.3	93.7
×8	**96.8**	**91.4**	**95.8**	**98.1**	**97.0**	**95.8**

[8] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2016: 770-778.

[9] Soto C, Yoo S. Visual detection with context for document layout analysis[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA: Association for Computational Linguistics, 2019: 3464-3470.

[10] Zhong X, Tang J, Yepes A J. Publaynet: largest dataset ever for document layout analysis[C]//2019 International Conference on Document Analysis and Recognition (ICDAR). Piscataway, NJ: IEEE, 2019: 1015-1022.

[11] Xu Y, Li M, Cui L, et al. LayoutLM: Pre-training of text and layout for document image understanding. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 1192-1200.

[12] Vaswani A , Shazeer N,Parmar N, et al. Attention is all you need [C]// Advances in Neural Information Processing Systems. 2017: 5998-6008.

[13] Wang X, Girshick R, Gupta A, et al. Non-local neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway, NJ: IEEE, 2018: 7794-7803.

[14] Cao Y, Xu J, Lin S, et al. Gcnet: Non-local networks meet squeeze-excitation networks and beyond[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Piscataway, NJ: IEEE, 2019: 1971-1980.

[15] Hu Han, Zhang Zheng, Xie Zhenda, et al. Local relation networks for image recognition[C]//In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2019: 3464–3473.

[16] Ramachandran P, Parmar N, Vaswani A, et al. Stand-Alone Self-Attention in Vision Models[C]//Neural Information Processing Systems. Cambridge: MIT Press. 2019: 68-80.

[17] Dosovitskiy A, Beyer L,Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale [EB/OL].(2021-01-03)[2022-11-12]. https :// arxiv. org/ pdf/2010. 11929.

[18] Touvron H , Cord M ,Douze M , et al. Training data-efficient image transformers & distillation through attention[C]//Proc of International Conference on Machine Learning. 2021: 10347-10357.

[19] Carion N , Massa F, Synnaeve G, et al. End-to-end object detectionwith transformers [C]//Proc of European Conference on ComputerVision. 2020: 213-229.

[20] Liu Ze, Lin Yutong, Cao Yue, et al. Swin transformer: hierarchical vi-sion transformer using shifted windows [C]//Proc of IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEEPress, 2021: 10012-10022.

[21] Ba J L, Kiros J R, Hinton GE. Layer normalization[EB/OL].(2016-07-21)[2022-11-12]. https ://arxiv. org/pdf/1607.06450.

[22] D. Hendrycks and K. Gimpel. Gaussian error linear units (GELUs)[EB/OL]. (2016-06-27)[2022-11-12]https ://arxiv. org/pdf/1606.08415.

[23] Gao Liangcai, Yi Xiaohan, Jiang Zhuoren, et al. ICDAR2017 competition on page object detection[C]//2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Washington DC: IEEE Computer Society, 2017: 1417-1422.

[24] He Kaiming, Gkioxari G, Dollar P, et al. Mask r-cnn[C]//2017 IEEE International Conference on Computer Vision (ICCV), Piscataway, NJ: IEEE Press, 2017: 2961-2969.

[25] Zhang Shifeng, Chi Cheng, Y. Yao Yongqiang, et al. Bridging the Gap Between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Piscataway, NJ: IEEEPress, 2020: 9756-9765.

[26] Li Kai, Wigington C, Tensmeyer C, et al. Cross-Domain Document Object Detection: Benchmark Suite and Method[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEEPress, 2020.

[27] Bolya D, Zhou Chong, F. Xiao Fanyi, et al. YOLACT++ Better Real-Time Instance Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(2): 1108-1121.

[28] Zhang Peng, Li Can, Qiao Liang, et al. VSR: A Unified Framework for Document Layout Analysis Combining Vision, Semantics and Relations[C]//2021 16th IAPR International Conference on Document Analysis and Recognition (ICDAR). Springer, Cham, 2021: 115-130.

[29] Li Xiaohui,Yin Fei, Liu Chenglin. Page Object Detection from PDF Document Images by Deep Structured Prediction and Supervised Clustering[C]// 2018 24th International Conference on Pattern Recognition (ICPR). Piscataway, NJ: IEEE, 2018: 3627-3632.

[30] Xu, Y. LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding, 2020. doi: 10.48550/arXiv.2012.14740.