YOLOLayout: Multi-Scale Cross Fusion Former for Document Layout Analysis

Zhangchi Gao [1] and Shoubin Li [1]
[1]The Institute of Software, Chinese Academy of Sciences, Beijing, China
shoubin@iscas.ac.cn

***Abstract*—Document layout analysis (DLA) is a technique used to locate and classify layout elements in a document, such as Table, Figure, List, and Text. While deep-learning-based methods in computer vision have shown excellent performance in detecting Text and Figures, they are still unsatisfactory in accurately recognizing the blocks of List, Title, and Table categories with limited data. To address this issue, we propose a single-stage DLA model that incorporates a Multi-Scale Shallow Visual Feature Enhancement Module (MS-SVFEM) and a Multi-Scale Cross-Feature Fusion Module (MS-CFF). The MS-SVFEM extracts multi-scale spatial information through the channel attention module, spatial attention module, and multi-branch convolution. The MS-CFF fuses different level features through an attention mechanism. The experiments showed that the mAP accuracy of YOLOLayout compared to the baseline model is 2.2% and 1.5% higher on the PubLayNet Dataset and the ISCAS-CLAD dataset.**

***Index Terms*—Document Layout Analysis, Document Object Detection, Document Structure**

I. Introduction

Document layout analysis (DLA) is a crucial research area that utilizes object detection or semantic segmentation techniques to delineate different regions within a document. DLA plays a vital role in various document-related applications, such as document understanding [1], knowledge extraction [2, 3], and optical character recognition (OCR). As fundamental research, DLA can significantly enhance the performance of other related tasks.

With the advancement of deep learning, both academia and industry have conducted extensive research on DLA. Although many deep-learning-based methods from computer vision have already achieved excellent performance in detectingFigure from documents, With a limited sample, they are still unsatisfactory in recognizing theList,Title andTable category blocks in DLA.

The analysis of the above issues is as follows: Firstly, theTitle occupies a relatively small area on the document page, making it harder for the model to accurately locate and recognize it. Secondly, the visual presentation of theList andText elements in the document is quite similar, leading to confusion about the model's ability to differentiate between the two. Finally, the complexity of the size and type ofTable, as well as the varied texture of different types of Tables, significantly increases the difficulty of the model's recognition process.

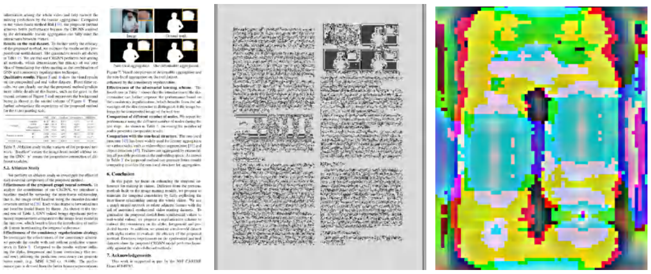

Fig. 1. A document feature map was created using the PubLayNet dataset. The left image shows the original sample, while the middle and right images display shallow and deep visual features, respectively. The shallow visual features were obtained from the first ConvBlock of the backbone network, while the deep visual features were obtained from the last ConvBlock.

The color and texture of document images differ from those of natural scenes, where document images typically have a single background and abundant element textures, and contain rich spatial and texture information in their shallow visual features. Existing methods primarily focus on learning from high-level channels while disregarding the knowledge in low-level channels. As demonstrated by [4], shallow visual features can improve object recognition in natural scenes. Figure 1 displays the shallow and deep visual features of the document image. The shallow visual features have distinct structured and texture characteristics, while the deep visual features have abstract semantic features.

Based on the above analysis, this paper proposes a method that integrates shallow visual features to improve the model's recognition effect onTable,List,Title under limited data conditions.

The contributions of our work can be summarized as follows:

- This paper presents a shallow visual feature enhancement module(MS-SVFEM) that utilizes spatial location information of layout blocks contained in shallow visual features to improve the recognition accuracy ofTitle.
- Propose a multi-scale fusion module(MS-CFF) that uses an attention mechanism to achieve adaptive fusion of deep and shallow visual features, improving the recognition accuracy of multi-scale layout blocks.
- An ISCAS-CLAD dataset for Chinese Document Layout Analysis has been released, which comprises 3000 training samples and 600 test samples.
- Our YOLOLayout model has been evaluated on the validation set provided by the dataset PubLayNet[5], with a 2.2% improvement in mAP compared to the

baseline model, while on ISCAS-CLAD, the mAP of YOLOLayout is 1.5% higher than that of the baseline model.

II. Related Work

Early DLA work can be divided into two categories, i.e., top-down and bottom-up strategies. Top-down approaches divide pages into text lines, words, etc. Representative works include texture-based analysis [6], run-length smearing [7], DLA projection profiling [8] and white space analysis [9]. The bottom-up approach [10–13] divides the objects into text lines and paragraphs by using the local characteristics of the object. With the rapid development of deep learning, some CNN [14] and Transformer [15] based methods have been proposed with impressive performance.

A. Document Layout Analysis

With the development of deep learning, many effective methods have been proposed and achieved good results in the field, and convolutional neural networks (CNNs) have become the main component of state-of-the-art DLA techniques. Most of the deep learning-based DLA methods are inspired by full convolutional neural networks (FCNs). [16] used FCNs with multi-scale features for document semantic segmentation. [17] adapted the full convolutional network (FCN) [18] to detect layout element within the page. [19] attempted DLA using a natural scene object detector. For more complex table data, [20] used Faster R-CNN to identify its structure and parse the content.

B. Multi-scale Feature Fusion in DLA

Multi-scale feature fusion is a basic method to solve the problems of large-scale variations of objects and complex scenes in visual detection tasks. Feature fusion work in natural scene target detection contains FPN [21], PANet [22], NAS-FPN [23], BiFPN [24], ASFF [25], etc. FPN integrates backbone features from different stages through a top-down path. Based on FPN, PANet enhances the whole hierarchy using top-down paths. BiFPN enables simple and fast multi-scale feature fusion through bi-directional cross-scale connections. In the field of DLA, there have been two recent advancements. Firstly, in a paper by [26], a dynamic residual fusion module was proposed to combine high-dimensional features with low-dimensional features. This approach successfully recovered image details while preserving category semantic information. Secondly, [27] developed a dynamic edge feature embedding block that combines learnable weights from different layers with edge features.

III. Methodology

This section provides an overview of YOLOLayout and its structural approach. Inspired by the work in [28, 29], we propose a Multi-Scale Cross Feature Fusion Module (MS-CFF) for global modeling of diff-level features and a Multi-Scale Shallow Visual Feature Enhancement Module (MS-SVFEN) to enhance the spatial information of shallow feature maps.

A. Overview

The network architecture is shown in Figure 2. We use the top-down and bottom-up paths of FPN [21] and PANet [22] as a base framework for feature fusion. In order to enrich the spatial location information in the high-level features, we use an MS-SVFEM to extract the spatial location information in the low-level features, and an MS-CFF to adaptively fuse the features of different levels. Among them, the MS-CFF is designed with the idea of ViTBlock [29].

B. Multi-Scale Shallow Visual Feature Enhance Module

Shallow visual features contain rich texture information and spatial location information. In this paper, we propose an MS-SVFEM to enrich the location information of a large-resolution detection head. The structure of MS-SVFEM is shown in Figure 3.

The input features are first enriched with texture features using a Channel Attention Module(CAM), and then the spatial information is reinforced using a Spatial Attention Module(SAM) to output a spatial attention map. The channel attention output is divided into three branches, each of which is processed through 3×3 convolution with different dilation rates to capture dependencies at different scales and learn more nonlinear features. The output of the three parallel convolutions is fused with the spatial attention map to obtain three spatially-informed enhanced features. These three features are then aggregated and processed through a 1×1 convolution to map them to a low-dimensional space. The goal of training is to extract independent features, so the aggregation of strongly correlated features speeds up convergence. Finally, using a 1×1 convolutional layer, adjust the feature channels to achieve enhanced shallow visual features.

C. Multi-Scale Cross Feature Fusion

FPN [23] proposes a practical fusion framework to solve the multi-scale problem by top-down path fusion features.PSPNet [30] uses pyramid pools to extract global context. However, both of them are used to recover small-scale feature maps to the original feature maps by bilinear upsampling in the absence of shallow visual features, which will cause the missing misalignment of the spatial location of the original large-scale feature maps.PANet [22] investigates an additional bottom-up path that uses shallow features at different scales to recover the boundary details gradually. However, combining different scale features may destroy the semantic category information and cause the wrong classification. Therefore, with the idea of MobileViT [29], we propose a multi-scale cross-feature fusion module (MS-CFF is shown in Figure 4) to fuse deep and shallow visual features.

In detail, given the input tensor$\underline{X}\in\underline{R_{H\times W\times C}}$ and$\underline{Y}\in\underline{R_{H\times W\times C}}$, MS-CFF applies an$\underline{n}\times\underline{n}$ standard convolution layer followed by Depth-Wise (or 1×1) Convolution to produce$\underline{X_L}\in\underline{R_{H\times W\times d}}$ and$\underline{Y_L}\in\underline{R_{H\times W\times d}}$ respectively.$\underline{n}\times\underline{n}$ convolution locally characterizes the input tensor, while Depth-Wise Convolution projects the tensor

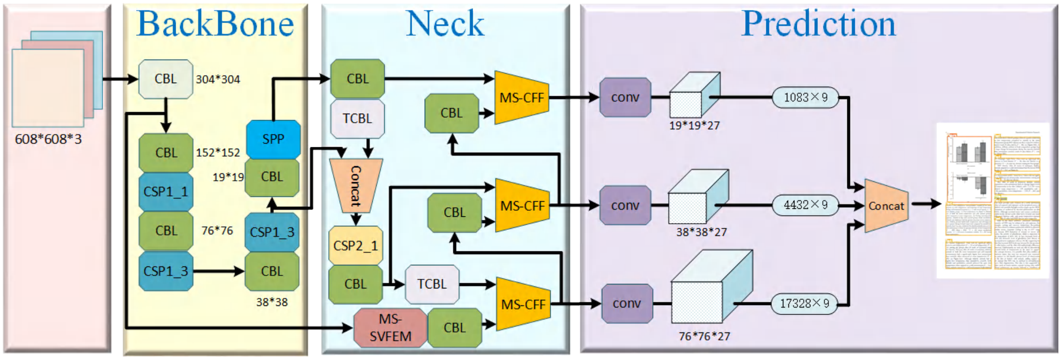

Fig. 2. The architecture of The YOLOLayout. It is divided into Backbone, Neck, and Prediction, three parts

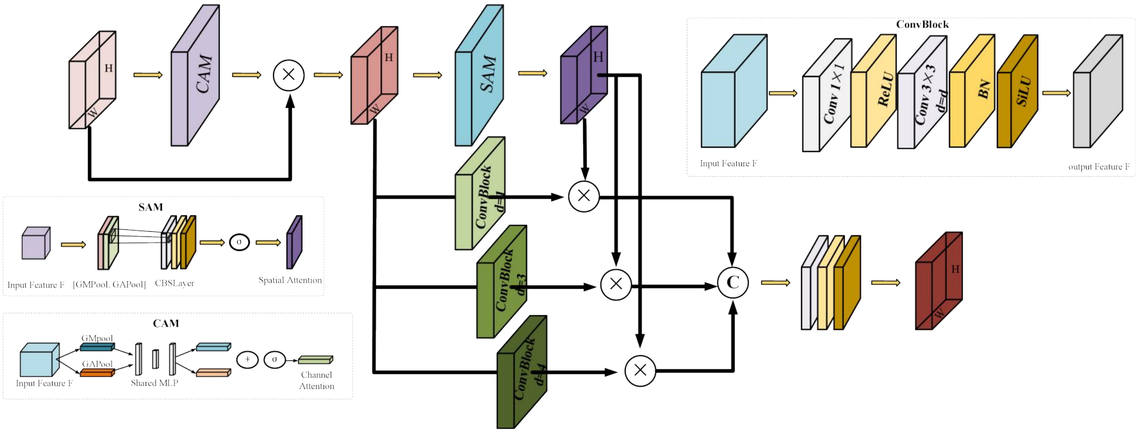

Fig. 3. The architecture of the MS-SVFEM. In the first stage, we refine the channel features using channel attention. In the second stage, we obtain the location features at different scales by using spatial attention and multi-branch convolution. Finally, in the third stage, we fuse the multi-branch to get the final output.

to the high-dimensional space ($d > C$), which enriches the channel information of the features.

To let the deep features with spatial inductive bias and shallow visual features do long-dependent matching,$X_L \in R_{H\times W\times d}$ and$Y_L \in R_{H\times W\times d}$ are divided into$N$ non-overlapping flat Patches$X_U \in R_{P\times N\times d}$ and$Y_U \in R_{P\times N\times d}$, where$p = wh$,$N = \frac{HW}{P}$ is the number of patches, and$h \leq n$ and$w \leq n$ are height and width of a patch respectively, and the sequences ofX_U andY_U are sent to Cross-transformer to get the relationship$X_G \in R_{P\times N\times d}$ of each Patches after fusion,where$X_G \in R_{P\times N\times d}$ as:

$$X_G(P) = Cross - Attention(X_U(P), Y_U(P)) \quad (1)$$

We unfold$X_G \in R_{P\times N\times d}$ to obtain$X_F \in R_{H\times W\times d}$.$X_F$ is then projected to a low$C$-dimensional space using a Depth-Wise convolution and combined with$X$ and$Y$ via Concatenation. Another$n \times n$ convolution layer is then to fuse these concatenated features.

D. Cross-Attention

Feature fusion refers to the process of combining features from different layers or branches by means of computations in order to address the issue of insufficient feature information. Typically, this technique is carried out using simple operations such asSUM or$CONCAT$, but these methods may fail to fully utilize the information contained within the features. To enhance feature fusion across different levels, a cross-attention(Figure 5) mechanism is employed to establish connections between different features. This can be represented as follows:

$$Head_i = Cross\!-\!Attention(XW_{Q_i}, YW_{K_jV_j}) \quad (2)$$

$$MHead(X,Y) = Cat(Head_1,..., Head_n)W_O \quad (3)$$

In detail, given the input tensor$X \in R_{B\times N\times C}$,$Y \in R_{B\times N\times C}$,$X$ are and$Y$ are projected to the$C$-dimensional space and$d$-dimensional space ($d = 2C$) by 1 × 1 Conv layer.the$C$-dimensional features are expressed as$Q$-vector and the$d$-dimensional features are divided intoKV-vector, This can be represented as follows:

$$Q_X, K_Y, V_Y = Conv(X), Split(Conv(Y)) \quad (4)$$

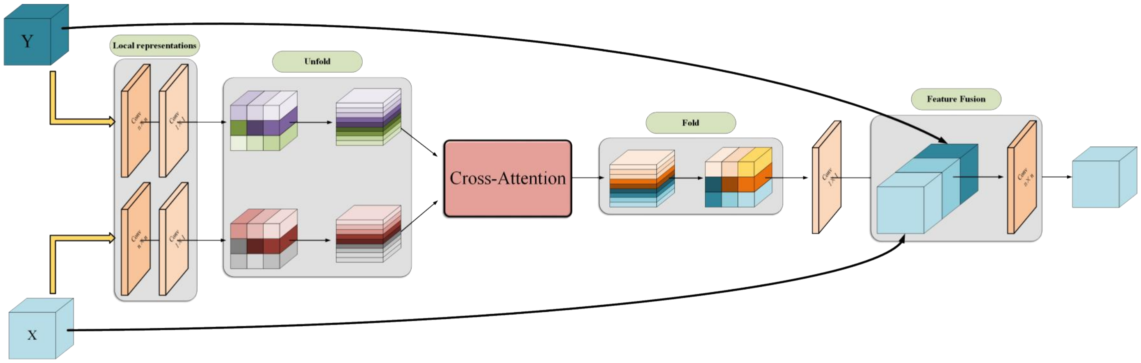

Fig. 4. The architecture of the MS-CFF. The orange and yellow blocks indicate 3×3 convolution and 1×1 convolution, which perform local representations of features. The dark blue block in the middle indicates the Cross-Attention module, which performs global representations of the unfolded patches.

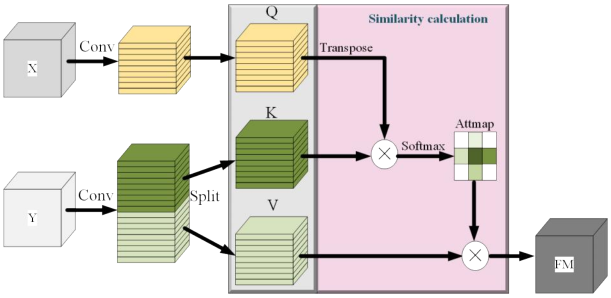

Fig. 5. The architecture of the Cross-Attention Module.

Table I: Sample Category Statistic of experimental dataset(PubLayNet).

Category	Training set	Validation set
Figure	178,252	88,625
Table	47,501	18,801
List	6,138	4,239
Text	7,729	4,769
Title	8,751	4,327
Total	248,371	120,761

TheQuery-vector of tensorX are matched with theKV pairs of tensorY to generate the connection strength between each patch. Then the similarity scores are calculated bysoftmax respectively, and finally, fused features are output:

$$F_M = Mul(Softmax(Q_X K_Y^T), V_Y) \tag{5}$$

IV. Experiment

A. Dataset

We validate the performance of the models on the PubLayNet [5] dataset and the ISCAS-CLAD [1] dataset. The PubLayNet dataset is automatically annotated by automatically matching the XML information of more than 1 million PDF articles publicly available on PubMed Central™ without manual annotation. The PubLayNet dataset contains five categories, Figure, Table, List, Text, and Title. A sample of the data is shown in Figure 6. The dataset contains a training set of over 360,000 document images and a validation set of 11,000 images.

The experiment in this paper is conducted on the complete PubLayNet-dev dataset (11,000 samples).To verify the model's ability to identify objects in limited data conditions, we refer to PubLayNet's five categories of distributions and randomly select a group of data according to a ratio of 7:3 for training and validation. Ultimately, the training set includes 25,000 document images, and the validation set includes 11,000 document images. The distribution of PubLayNet data is shown in Table I.

The ISCAS-CLAD dataset in this paper has 3000 training samples and 600 test samples, which are constructed by model pre-labeling and manual correction. To ensure the correctness of sample labeling, a senior researcher(leader) in the DLA and two Ph.D. candidates (members) were selected to form a data correction team. The leader explained the annotation rules to the members and provides assistance in the data correction process. After manual correction, the leader reviews and corrects the data again. The dataset contains ten categories, Text, Title, Figure, Figure caption, Table, Table caption, Header, Footer, Reference, and Equation. A sample of the data is shown in Figure 7. Data distribution is shown in Table II.

B. Implementation Settings

This section describes the experimental settings of this paper. The performance of our proposed YOLOLayout model is first compared with the baseline model YOLOV5 on the PubLayNet dataset to observe the effectiveness of DLA to fuse shallow visual features. In addition, we also choose pre-trained Mask RCNN and Faster RCNN on PubLayNet for comparison with YOLOLayout. Since we want to compare the localization ability of YOLOLayout on small DLA datasets, we reproduce the baseline based on their experimental settings. Finally, we choose the MAP @ IOU [0.50:0.95] evaluation metric applied

[1] https://github.com/ISCAS-ITECHS/ISCAS-CLAD

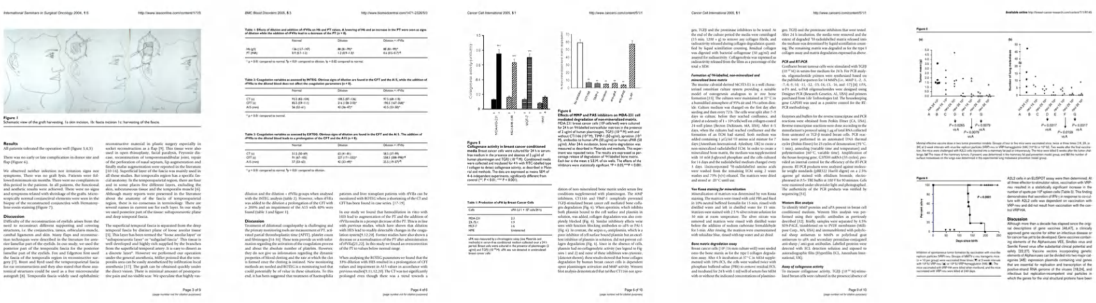

Fig. 6. The data example of PubLayNet.

Fig. 7. The data example of ISCAS-CLA

Table II: Sample Category Statistic of Chinese Layout Analysis dataset.

Category	Training set	Validation set
Text	15,072	1,859
Title	5,856	835
Figure	2,921	297
Figure caption	2,880	267
Table	758	140
Table caption	729	129
Header	7,596	1,064
Footer	2,304	237
Reference	1,846	275
Equation	1,220	120
Total	41,182	5,223

in the PubLayNet dataset paper as the evaluation metric for all our experiments.

Our YOLOLayout model is implemented based on the PyTorch framework. We use the official YOLOV5m pre-training weights provided by YOLOV5 as our initial weights. The model input resolution is set to 640×640, Adamw is used as the optimizer, the batch size is set to 32, the model is trained for a total of 150 epochs, and smooth L1 and Focal loss are used as the loss functions for object localization and classification. In MS-SVFEM, we use the output of the first layer of Conv after the Focus structure of the backbone network as the input of the MS-SVFEM module, where the input tensor size is 320×320×128 and the output tensor size is 160×160×256. In the MS-CFF module, we use this module in the PANet path for the fusion of deep and shallow visual features. Firstly, we downsample the shallow visual features to get the exact size resolution feature map as the deep features. We feed the two equal-size feature maps into the MS-CFF module for fusion, and finally, the feature maps output by the MS-CFF module are localized and classified.

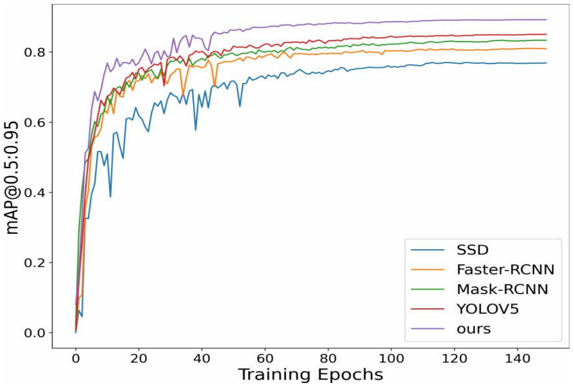

Fig. 8. Convergence curve shows the training accuracy curves of Mask-RCNN, Faster-RCNN, SSD, and our model on PubLayNet.

C. Results and Analysis

To verify the recognition capability of the YOLOLayout proposed in this paper on small and medium-sized DLA datasets, ablation experiments, stability experiments, and comparison experiments are conducted in the paper. (1) Ablation experiments, comparing the YOLOLayout model with the baseline model YOLOV5 on the PubLayNet dataset, to examine the effect of each module. (2) Stability experiments, were conducted on ISCAS-CLA to verify the stability of the YOLOLayout model. In the five-fold cross-validation experiments, ISCAS-CLA is divided into five equal parts, one part is

Table III: The abaltion experimental results of mAP@IOU0.50:0.95 of our method.

Section	YOLOV5	cYOLOLayout$_{MS-SVFEM}$	cYOLOLayout$_{MS-CFF}$	YOLOLayout$_{MS-SVFEM+MS-CFF}$
Figure	0.934	0.922	0.946	**0.949**
Table	0.924	0.928	**0.947**	0.945
List	0.785	0.82	**0.851**	0.846
Text	0.914	0.92	**0.928**	0.924
Title	0.726	**0.747**	0.719	0.731
Average	0.857	0.867	0.878	**0.879**

randomly taken out as the test set each time, and the remaining four parts are used as the training set, and five randomized trials are conducted in this way. (3) comparison experiments, we select Mask-RCNN and Faster-RCNN that have been pre-trained on PubLayNet and re-trained on a small randomly selected PubLayNet dataset to derive results for comparison with the performance of the YOLOLayout model.

a) Ablation experiments: Table III shows the accuracy of MAP@IOU[0.50:0.95] for the YOLOLayout model and all ablation experiments, where MS-SVFEM denotes the Multi-Scale Shallow Visual Feature Enhancement Module and MS-CFF denotes the Multi-Scale Cross Feature Fusion Module. Meanwhile, YOLOLayout$_{MS-SVFEM+MS-CFF}$ denotes our proposed YOLOLayout model, YOLOLayout$_{MS-SFVEM}$ denotes the YOLOLayout model without the MS-CFF, and YOLOLayout$_{MS-CFF}$ denotes the YOLOLayout model without the MS-SVFEM. We first conducted experiments with YOLOV5 as the baseline model, and as shown in Table III, the baseline model achieved 85.7% mAP. We further evaluate the impact of the Multi-Scale Cross Feature Fusion Module (MS-CFF) by attaching it to the detection head. Essentially, MS-CFF is designed to reduce the corruption of features by early fusion by aggregating different scale features with similarity calculations. Table III shows that the addition of the MS-CFF module yields 87.8% mAP, which is a 2.1% increase over the baseline model, and achieves the best accuracy in the detection of List, Table, and Text. In addition, the MS-SVFEM module is added to the baseline model, which brings a 1.0% improvement in mAP and achieves the best detection accuracy on the detection of Title. Finally, the YOLOLayout structure improves the mAP by 2.2% compared to the baseline model.

In addition, there are still some shortcomings in our proposed method. The first problem, although the model has improved the detection of Title after adding shallow visual features, the accuracy is low compared to other classes of document objects. The main reason for the inaccurate detection of titles is that the size of most Titles is smaller compared to the size of Text, Figure, List, and Table, which occupy fewer effective pixels in the document image and belong to small document targets, and even some titles cannot be accurately identified by human eyes. The second problem is that although the YOLOLayout structure achieves the best accuracy in the whole ablation experiment, it only achieves the best accuracy in the detection of Figure, and the total average mAP is only 0.1% higher compared to MS-CFF.

Table IV: The abaltion experimental results of mAP@IOU0.50:0.95 of our method.

Section	Faster-RCNN	Mask-RCNN	YOLOLayout
Text	0.910	**0.914**	0.909
Title	0.742	0.758	**0.789**
Figure	0.849	0.857	**0.867**
Figure caption	0.834	0.843	0.843
Table	0.851	0.871	**0.912**
Table caption	0.847	0.856	0.886
Header	0.759	0.761	0.778
Footer	0.593	0.601	0.612
Reference	0.908	0.904	0.901
Equation	0.767	0.764	0.775
Average	0.806	0.813	**0.828**

b) Stability experiments: Stability experiments are conducted on our dataset for YOLOLayout, Faster-RCNN, and MaskRCNN. When stability check experiments are performed on ISCAS-CLA, as shown in Table IV, the mAP accuracy of YOLOLayout for recognizing Title and Table is significantly better than Mask-RCNN and Faster-RCNN, and the total average accuracy is 2.2% higher than Faster-RCNN and 1.5% higher than Mask-RCNN. Among them, YOLOLayout recognizes Title and Table 5% and 6% better than Faster-RCNN, and 3% and 4% better than Mask-RCNN. These experimental results show that the YOLOLayout model proposed in this paper outperforms Faster-RCNN and Mask-RCNN on small layout analysis datasets and has good stability.

c) Comparisons experiments: To evaluate the performance of our models on small datasets, we add the mean Average Precision(mAP) of Mask-RCNN, Faster-RCNN, and SSD on the PubLayNet dataset as well as the accuracy of each document object class (Figure, Table, List, Text, Title). Where MaskRCNN and FasterRCNN are pre-trained models on PubLayNet provided by LayoutParser and use the training parameters provided by LayoutParser. We replace the convolutional mention neural network in SSD with a separable convolutional neural network and add the feature pyramid structure.

As shown in Table III, it can be seen that our results are better than the results of other models. It is worth noting that our YOLOLayout model outperforms Mask-RCNN, Faster-RCNN, and SSD in detecting Figure, Table, and List, indicating that shallow visual features contain more recognition

Table V: Comparisons with Prior Arts.

Model	Resolution	Figure	Table	List	Text	Title	Average
SSD [31]	640×640	0.878	0.882	0.690	0.840	0.589	0.774
YOLOV5	640×640	0.934	0.924	0.785	0.914	0.721	0.857
Faster R-CNN [32]	960×960	0.907	0.891	0.795	0.923	0.712	0.845
Mask R-CNN [33]	960×960	0.918	0.901	0.801	**0.931**	0.727	0.852
YOLOLayout(our)	640×640	**0.949**	**0.945**	**0.846**	0.924	**0.731**	**0.879**

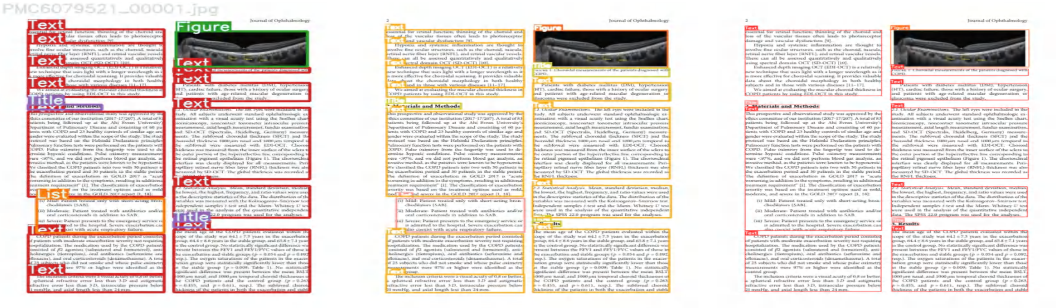

Fig. 9. result analysis on PubLayNet dataset. the first, second, and third columns represent ground truth, our proposed YOLOLayout result, and Mask-RCNN result, respectively.

information of structured regions; therefore, shallow visual features are effective for the model to detect complex structured document objects. The detection performance of Text and Title is only slightly lower than the detection performance of Text and Title of Mask-RCNN. Overall, YOLOLayout achieves state-of-the-art as seen in Table III compared with our previous work. Although the detection accuracy of List has been significantly improved, its detection and that of Title are still challenging.

We also compare the training convergence speed of the models in Table III on PubLayNet. As shown in Figure 8, our model has a competitive convergence speed and better detection performance.

d) Result Analysis.: The detection results obtained by YOLOLayout have been displayed in Figure 9. The left column shows the document image with Ground Truth, the middle column shows the document image predicted by YOLOLayout, and the right column shows the document image predicted by Mask-RCNN. We can observe that the predicted element positions in the YOLOLayout are correct (Figure IV-C0a) and that Mask-RCNN's prediction of Text and List is confused, misidentifying List as Text(Figure IV-C0a).

V. Conclusion

This paper presents a novel solution for constructing a generic DLA model. We explore the use of shallow visual features in the backbone network to enhance the expressiveness of the DLA model. We propose an MS-SVFEM that establishes dependencies of different lengths on the location information in the shallow features and can adaptively incorporate features of different scales. We use an MS-CFF to fuse deep features with shallow features, allowing the model to establish dependencies between semantic and spatial information. Experimental results show that our proposed YOLOLayout model exhibits excellent performance on the PubLayNet dataset. As a fundamental study, DLA can be applied to various fields. Our proposed YOLOLayout model has been applied to parsing PDF documents into HTML.

With the rapid development of deep learning in recent years, more and more excellent models have emerged, but most are based on high-quality data. Most of the public datasets for DLA tasks are English data, which can only meet the needs of some applications, so document data enhancement in my field is our next research direction.

References

[1] L. Ding, A. Goshtasby, On the canny edge detector, Pattern Recognition 34 (3) (2001) 721–725.

[2] J. Kittler, On the accuracy of the sobel edge detector, Image and Vision Computing 1 (1) (1983) 37–42.

[3] Y. Soullard, P. Tranouez, C. Chatelain, S. Nicolas, T. Paquet, Multi-scale gated fully convolutional densenets for semantic labeling of historical newspaper images, Pattern recognition letters (131-Mar.).

[4] J.-S. Lim, M. Astrid, H.-J. Yoon, S.-I. Lee, Small object detection using context and attention, in: 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2021, pp. 181–186. doi:10.1109/ICAIIC51459.2021.9415217.

[5] X. Zhong, J. Tang, A. J. Yepes, Publaynet: largest dataset ever for document layout analysis, in: 2019 International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2019, pp. 1015–1022.

[6] A. Asi, R. Cohen, K. Kedem, J. El-Sana, Simplifying the reading of historical manuscripts, in: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2015, pp. 826–830.

[7] W. Swaileh, K. A. Mohand, T. Paquet, Multi-script iterative steerable directional filtering for handwritten text line extraction, in: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2015, pp. 1241–1245.

[8] F. Shafait, T. M. Breuel, The effect of border noise on the performance of projection-based page segmentation methods, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (4) (2010) 846–851.

[9] F. Shafait, J. Van Beusekom, D. Keysers, T. M. Breuel, Background variability modeling for statistical layout analysis, in: 2008 19th International Conference on Pattern Recognition, IEEE, 2008, pp. 1–4.

[10] T. A. Tran, I.-S. Na, S.-H. Kim, Hybrid page segmentation using multilevel homogeneity structure, in: Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication, 2015, pp. 1–6.

[11] M. Mehri, P. Héroux, P. Gomez-Krämer, R. Mullot, Texture feature benchmarking and evaluation for historical document image analysis, International Journal on Document Analysis and Recognition (IJDAR) 20 (1) (2017) 1–35.

[12] Y. Lu, C. L. Tan, Constructing area voronoi diagram in document images, in: Eighth International Conference on Document Analysis and Recognition (ICDAR'05), IEEE, 2005, pp. 342–346.

[13] N. Vasilopoulos, E. Kavallieratou, Complex layout analysis based on contour classification and morphological operations, in: Proceedings of the 9th Hellenic Conference on Artificial Intelligence, 2016, pp. 1–4.

[14] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural computation 1 (4) (1989) 541–551.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30.

[16] D. He, S. Cohen, B. Price, D. Kifer, C. L. Giles, Multi-scale multi-task fcn for semantic page segmentation and table detection, in: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017.

[17] S. A. Oliveira, B. Seguin, F. Kaplan, dhsegment: A generic deep-learning approach for document segmentation, in: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2018.

[18] Long, Jonathan, Shelhamer, Evan, Darrell, Trevor, Fully convolutional networks for semantic segmentation, IEEE Transactions on Pattern Analysis & Machine Intelligence.

[19] Y. Xu, F. Yin, Z. Zhang, C.-L. Liu, et al., Multi-task layout analysis for historical handwritten documents using fully convolutional networks., in: IJCAI, 2018, pp. 1057–1063.

[20] S. Schreiber, S. Agne, I. Wolf, A. Dengel, S. Ahmed, Deepdesrt: Deep learning for detection and structure recognition of tables in document images, in: 2017 14th IAPR international conference on document analysis and recognition (ICDAR), Vol. 1, IEEE, 2017, pp. 1162–1167.

[21] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.

[22] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8759–8768.

[23] G. Ghiasi, T.-Y. Lin, Q. V. Le, Nas-fpn: Learning scalable feature pyramid architecture for object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 7036–7045.

[24] M. Tan, R. Pang, Q. V. Le, Efficientdet: Scalable and efficient object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10781–10790.

[25] S. Liu, D. Huang, Y. Wang, Learning spatial fusion for single-shot object detection, arXiv preprint arXiv:1911.09516.

[26] X. Wu, Z. Hu, X. Du, J. Yang, L. He, Document layout analysis via dynamic residual feature fusion.

[27] X. Wu, Y. Zheng, T. Ma, H. Ye, L. He, Document image layout analysis via explicit edge embedding network, Information Sciences 577 (2021) 436–448.

[28] S. Li, X. Ma, S. Pan, J. Hu, L. Shi, Q. Wang, Vtlayout: Fusion of visual and text features for document layout analysis, in: PRICAI 2021: Trends in Artificial Intelligence, Springer International Publishing, Cham, 2021, pp. 308–322.

[29] S. Mehta, M. Rastegari, Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer.

[30] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2881–2890.

[31] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in: European conference on computer vision, Springer, 2016, pp. 21–37.

[32] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: NIPS, 2016.

[33] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.