# Document Image Layout Analysis via MASK Constraint

**Jun He[1], Hanjie Zheng[2], Tianlong Ma[1,*]**
*[1]East China Normal University, Shanghai, China*
*[2]Vernon Secondary School, Vernon, Canada*
**Corresponding Author.*

**Abstract: Document layout analysis plays an essential role in computer vision. With the development of deep learning, more and more deep learning methods are proposed to solve some challenges in document layout analysis. Semantic segmentation-based and object detection-based methods are two mainstream approaches for document layout analysis. Compared with methods based on semantic segmentation, methods based on target detection have certain advantages in ensuring the integrity of target objects, especially with the proposal of Mask R-CNN. However, since the document layout analysis task is different from the general target detection task, there is a particular semantic gap in the document layout analysis (i.e the image to be detected may contain text), and the Mask R-CNN cannot solve this problem well. Therefore, we design a hierarchical information augmentation module, which can fully utilize low-dimensional detail information and maintain high-dimensional semantic information. In addition, we propose a novel MASK-constrained module, which ensures that the global semantic information of the input module can be further mined by embedding MASK information in the input image. Furthermore, to combat the issue of overlapping bounding boxes arising from Mask R-CNN processing, we propose a Constrained Aggregation method. Finally, we validate our approach using benchmark datasets featuring complex layouts (such as DSSE-200 and FPD). The results underscore the significant performance gains achievable with our proposed method.**

**Keywords: Document Layout Analysis; Computer Vision; Semantic Segmentation; Object Detection**

## 1. Introduction

The task of document layout analysis occupies a crucial role within the field of computer vision, serving as a fundamental process for understanding and interpreting the structured components of document images. The primary objective of this analysis is to employ advanced techniques, notably semantic segmentation, to systematically categorize and separate high-order semantic regions within these images. These regions typically include elements such as figures, tables, text, and the background. The advent and progressive development of deep learning technologies have significantly enhanced the efficacy of document layout analysis. This is particularly evident in its application to documents that do not adhere to the traditional Manhattan layout, which is characterized by rectangular configurations or other types of flow layouts. Thanks to these advancements, contemporary deep detection models have demonstrated remarkable capabilities in identifying and analyzing document layouts, achieving notable success. This progress is well-documented in the literature, with several studies [1–4], highlighting the substantial advancements and performance enhancements in the field of document layout analysis enabled by deep learning methodologies.

From a technological perspective, the methods employed in document layout analysis predominantly revolve around semantic segmentation frameworks and detection-based frameworks, each with its distinct methodologies and benefits. Detection-based approaches, in particular, offer significant advantages in maintaining the integrity of segmentation outcomes. Notably, the most acclaimed achievements within this domain have leveraged the capabilities of Mask R-CNN, as evidenced by pioneering research such as that conducted by [5,6]. These studies have demonstrated the effectiveness of Mask R-CNN in the context of deep layout analysis, marking a significant milestone in the field. Despite these advancements, the application of Mask R-CNN to document layout analysis is not without its challenges, especially when addressing documents that deviate from the conventional Manhattan layout,characterized by their non-rectangular structures. The inherent limitation of Mask R-CNN, which typically generates detection boxes in rectangular shapes, poses specific obstacles inaccurately detecting non-Manhattan layouts. This issue is illustrated in Figuer 1, which showcases the semantic discrepancies that can arise when Mask R-CNN, retrained on a dataset specific to document layout analysis, is applied. These discrepancies can mislead the model, leading to incomplete or inaccurate detection outcomes. The root of these challenges lies in the unique characteristics of the documents themselves. Many documents incorporate images that contain substantial amounts of text, yet these texts are categorized separately from the images. This distinction introduces a semantic gap in classification, culminating in the truncation of some semantic objects. Such challenges underscore the complexity of document layout analysis and the need for refined approaches that can adeptly navigate the nuanced distinctions within document compositions.

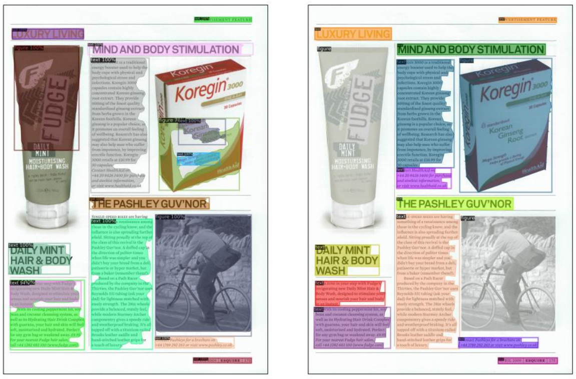
LUXURY LIVING
MIND AND BODY STIMULATION
Koregin
FUDGE
THE PASHLEY GUV'NOR
DAILY MINT HAIR & BODY WASH
LUXURY LIVING
MIND AND BODY STIMULATION
Koregin
FUDGE
THE PASHLEY GUV'NOR
DAILY MINT HAIR & BODY WASH

**Figure 1. Left: Layout Detected Using Mask R-CNN; Right: the Ground-True of the Layouts. Mask R-CNN Needs to be Trained on the Target Dataset.**

In the realm of document layout analysis, the introduction of E3Net [7] marked a significant advancement by incorporating edge information to delineate document boundaries more effectively. Inspired by the innovations brought forth by E3Net, our approach seeks to integrate a structure imbued with global semantic information to address the issue of semantic discontinuities. To this end, we propose a novel methodology that involves embedding the MASK of an image as a means to preserve and emphasize global information within the document. This technique is aimed at mitigating the challenges associated with semantic gaps, thereby enhancing the accuracy and integrity of document layout analysis. Building upon this foundation, we introduce the dual information augmentation module, a concept designed to augment and enrich contextual information within the document. This module distinguishes itself from E3Net through its unique approach to processing fused multi-channel data. We have developed a lightweight structure that draws inspiration from the principles of Spatial Pyramid Pooling, as detailed by [8]. This structure is tailored to efficiently manage and analyze the complexities of multi-channel data, facilitating a more nuanced and comprehensive understanding of document layouts.

Moreover, as depicted in Figure. 1 (bottom left), the utilization of Mask R-CNN for document layout analysis frequently results in the occurrence of overlapping detection regions. This phenomenon complicates the interpretation and classification of document elements, necessitating a robust solution to effectively address these overlaps. In response to this challenge, we have developed a post-processing algorithm that borrows concepts from clustering methodologies. This innovative approach begins by selecting a random point within the document that does not belong to the background class. Following this initial step, the algorithm examines the four neighboring points surrounding the selected point. If these neighbors exhibit similar characteristics—determined based on their RGB values—they are merged into a single region. This merging process continues until a point is reached where the four neighboring points are no longer similar. The algorithm then proceeds to identify another unvisited, random point that does not fall under the background classification, repeating the aforementioned steps until a state of convergence is achieved. This method proves effective in resolving the issue of overlapping detection regions, presenting a significant improvement in the handling of document layouts, especially those that are more complex than the standard Manhattan layout. It is particularly noteworthy that our proposed MASK constraint module demonstrates enhanced performance when applied to documents featuring intricate layouts. These non-Manhattan layouts, which often deviate from simple rectangular or traditional Manhattan configurations, present unique challenges that our method is specially designed to overcome. By implementing MASK constraints, we aim to significantly enhance the accuracy and efficiency of document layout analysis, particularly for documents with complex structural characteristics.

Our contributions are summarized as follows.

To address the semantic gap in current detection-based document layout analysis, we propose the dual information augmentation module. This module enhances the process with contextual information, effectively bridging the gap. By doing so, it improves the accuracy of interpreting complex document elements, providing a more detailed and precise analysis.

We introduce the MASK constraint module to capture global information often overlooked by Mask R-CNN-based methods. Alongside, we present a post-processing algorithm, Constrained Aggregation, to address the issue of overlapping detection boxes.

Our proposed methods have significant performance for documents with non-Manhattan layouts and are suitable for dealing with complex layouts. We conduct experiments on DSSE and FPD datasets to demonstrate the superiority of our proposed method.

## 2. Methodology

### 2.1 Overview

Illustrated within the framework depicted in Figure 2, we have developed a document layout analysis framework, termed MCNET, which is founded upon the principles of Mask R-CNN. Utilizing Mask R-CNN as its core architecture, MCNET integrates a Hierarchical Information Augmentation module specifically engineered to bolster contextual information throughout the document analysis process. In tandem with this augmentation module, we introduce a MASK constraint mechanism. This innovative approach is designed to fully interpret and leverage the information presented in a 4-channel input format. By doing so, it ensures the effective integration of comprehensive global information into the analysis, thereby enhancing the framework's capability to accurately and efficiently parse document layouts. This strategic incorporation of hierarchical information augmentation alongside the MASK constraint underscores MCNET's advanced methodology in addressing the complexities inherent in document layout analysis, facilitating a more nuanced and detailed examination of document structures.

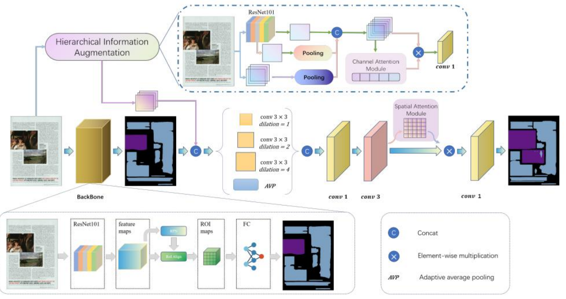

**Figure 2. The Architecture of the MCNet the Backbone is Mask-RCNN.**

### 2.2. Hierarchical Information Augmentation

As outlined in Section 1, Mask R-CNN encounters limitations in fully leveraging appearance and semantic information, leading to the omission of crucial contextual details. In response to this challenge, we introduce a hierarchical information augmentation module designed to enhance the utilization of both appearance and semantic data effectively.

Initially, we extract two key outputs from ResNet101: the middle state and the final output. These represent shallow, fine, detailed appearance information and deep, coarse, semantic information, respectively. Following this extraction, a pooling operation is applied to downsample these layers. We then concatenate them with the RGB data of the image, resulting in a composite 5-channel object. To further refine this object, a channel attention mechanism is employed. This mechanism is crucial for deriving channel-level attention weights, which are used to transform the initial 5-channel object into a 3-channel enhanced image.

Subsequently, we integrate the output of Mask R-CNN, which contains fine semantic information, with the previously enhanced image data. This integration process results in the creation of a hierarchically augmented 4-channel input. This innovative approach allows for a more comprehensive inclusion of both detailed appearance and semantic information,thereby addressing the initial shortcomings identified with Mask R-CNN's handling of contextual information. Through this hierarchical augmentation, we aim to significantly improve the accuracy and effectiveness of document layout analysis by enriching the contextual depth available for processing.

### 2.3. MASK Constraint

In our approach, we have devised a pyramid pool model that mirrors the concept of utilizing a 4-channel input more exhaustively. This model is specifically designed to capture the essence of information from various receptive fields, ensuring that the output from each convolution layer consists of eight channels. This strategy enables a comprehensive analysis of the input data, catering to the diverse dimensions and scales inherent within document images. To further enhance the model's capability in leveraging global information, we have integrated the technique of adaptive average pooling. This method is instrumental in extracting global information across the entire object, providing a holistic view of the document layout. Following this step, we concatenate the extracted 32- channel feature information, allowing for a more complex and informative representation of the document's layout. Subsequently, this concatenated feature set transforms a 1 × 1 convolution operation, which serves to adjust the number of channels to eight. This is followed by a 3 × 3 convolution, which maintains the channel count at eight. At this juncture, we introduce a spatial self-attention mechanism. This addition is pivotal in hierarchically extracting spatial information, enabling the model to discern and prioritize different areas within the document based on their relevance and content. The final stage of our model employs another 1 × 1 convolution, which is tasked with refining the channel count to six, corresponding to the number of categories identified within the document layout analysis. This step ensures that the model's output is tailored to the specific classification needs of the task, facilitating accurate and efficient categorization of document elements. Through this meticulously structured process, our model aims to provide a nuanced and detailed analysis of document layouts, accommodating the complex and varied structures they may present.

### 2.4. Constrained Aggregation

Mask R-CNN operates on a detection-based approach, setting it apart from semantic segmentation techniques primarily due to the characteristic overlap in detection results. Such overlaps can often introduce ambiguity into the interpreted outcomes, complicating the process of accurately delineating document layouts. To address this challenge, we introduce a post-processing algorithm we have named "Constrained Aggregation." The methodology behind Constrained Aggregation begins with aggregating all coordinate points from the prediction results that are not identified as part of the background into a comprehensive list. Following this initial step, the algorithm selects an arbitrary point from the list and initiates a search for similar points within its immediate four neighborhoods, employing RGB numerical operations for comparison. Should similar points be detected, the search extends to these points, subsequently removing them from the list to prevent redundancy. This process is repeated until no further similar points are found, at which point the algorithm proceeds to the next point on the list. This iterative process continues until the list is fully exhausted, culminating in the amalgamation of identified areas into coherent segments. This post-processing strategy, Constrained Aggregation, is meticulously designed to overcome the inherent limitations of detection-based methods like Mask RCNN, especially the issue of overlapping detection results. By methodically consolidating similar points and effectively organizing them into distinct segments, Constrained Aggregation enhances the clarity and precision of the detection outcomes, significantly reducing the ambiguity caused by overlaps and contributing to a more accurate interpretation of document layouts.

## 3. Experiment

In our experiments, we assess the performance of MCNet through a comprehensive evaluation framework that incorporates several key metrics, including accuracy, precision, recall, and the F1 score. These metrics collectively provide a holistic view of the model's effectiveness in document layout analysis.

### 3.1. Datasets

For the evaluation of MCNet, we leverage two benchmark datasets known for their relevance and complexity in the domain of document layout analysis.

DSSE-200: Introduced by [1], the DSSE-200 dataset stands out as one of the largest and most complex datasets available for document layout analysis. It is meticulously manually annotated and encompasses a wide array of document images. This diversity includes presentations (PPT), vintage newspapers, books, magazines, and academic papers, presenting a comprehensive challenge that spans various types and formats of documents.

FPD: Developed by [9], the FPD dataset represents the pioneering effort in manual annotation for fine-grained segmentation within document layout analysis. Comprising 66 documents characterized by non-Manhattan layouts, the FPD dataset is distinguished by its focus on intricate segmentation challenges. These documents are sourced from pages of sophisticated Chinese and English magazines, featuring a variable page size that adds a layer of complexity to the task at hand.

**Table 1. Overall Performance of MCNet**

| Method | Acc | P | R | F1 |
|---|---|---|---|---|
| DSSE-200 | 0.92 | 0.88 | 0.85 | 0.81 |
| FPD | 0.90 | 0.92 | 0.88 | 0.89 |

### 3.2. Qualitative Results

**DSSE-200.** As indicated in Table 1, our approach to Document Semantic Structure Extraction (DSSE) achieves an impressive accuracy rate of 92%. Furthermore, we present the prediction outcomes in Figure. 3-Left, where it is evident that our results closely approximate the true values. This accuracy is particularly noteworthy in terms of preserving the integrity of regions within the document. Our method demonstrates superior performance compared to conventional semantic segmentation techniques, highlighting its effectiveness inaccurately identifying and extracting semantic structures within documents. This achievement underscores the potential of our approach in advancing the field of document analysis by providing reliable and precise extraction of document semantic structures.

**FPD.** FPD is a dataset characterized by its complex annotations, setting it apart from simpler datasets such as DSSE-200 due to its increased intricacy and detail. Notably, FPD focuses on the annotation of non-Manhattan layouts, presenting a more challenging scenario for document layout analysis. In our research, we employed MCNet to perform experiments on this dataset, with the experimental outcomes detailed in Table 1. Furthermore, the prediction results are visually depicted in Figure 4-Right.

The comparative analysis reveals that the completeness of the region and the precision of the recognition results achieved by our method bears a close resemblance to those observed in the actual image. This similarity is particularly pronounced after the implementation of the Mask constraint, which significantly enhances the completeness of the analysis. Moreover, the application of the Mask constraint effectively minimizes the occurrence of splitting issues that typically arise due to semantic gaps. This improvement underscores the efficacy of our approach in dealing with complex document layouts, especially those that deviate from conventional Manhattan configurations, thereby highlighting the potential of our methodology in advancing the field of document layout analysis.

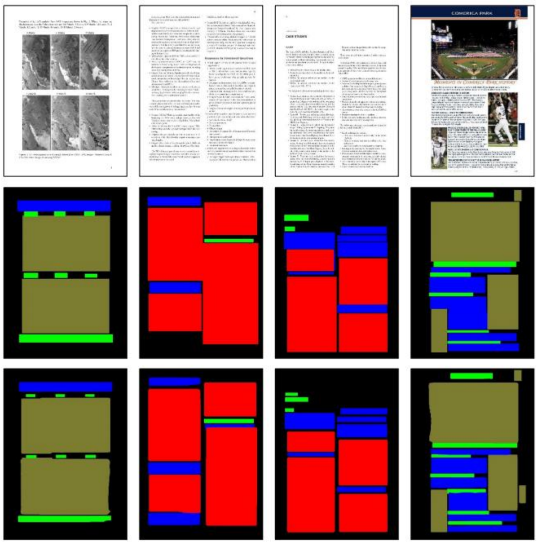

**Figure 3. Example Real Documents and Their Corresponding Segmentation on DSSE-200. Top: Original. Middle: Ground-Truth. Bottom: Predictions. Segmentation Label Colors are: Text, Title, List, Table, Figure, Background [DSSE-200].**

### 3.3. Comparisons with Prior Arts

In alignment with the experimental framework established by [1], we conducted a series of tests using the DSSE-200 dataset. The outcomes of our experiments, as detailed in Table 2 and Table 3, demonstrate that our model outperforms the results obtained in the study by [1]. Notably, our model exhibits superior performance, particularly in categories that impose total constraints. Furthermore, we undertook a comprehensive review and replication of several classic methodologies within this domain. Through this comparative analysis, it became evident that our approach not only holds distinct advantages over the benchmarks set by [1] but also demonstrates competitive strengths when measured against a spectrum of traditional methods. This indicates the robustness and effectiveness of our model in handling complex document layout analysis tasks, underscoring its potential for broader application in the field.

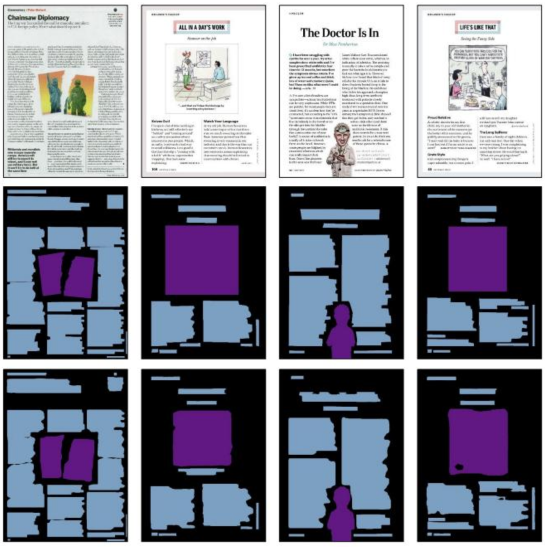

**Figure 4. Example Real Documents and Their Corresponding Segmentation on FPD. Top: Original. Middle: Ground-Truth. Bottom: Predictions. Segmentation Label Colors are: Text, Figure, Background**

### 3.4. Ablation Study

In this section, we evaluate the MCNet’s performance on the DSSE-200. We present quantitative and qualitative results that show the usefulness of the MCNet model.

**Model architecture.** In this section, our objective is to elucidate the architectural advantages of our model, particularly emphasizing the rationale behind the selection of only four channels as the input for our Mask Constraint module.

**Table 2. Per-Category Comparison Based on IoU Scores (%) on DSSE-200.**

| Method | figure | table | section | caption | list | para. | mean |
|---|---|---|---|---|---|---|---|
| MFCN [1] | 83.7 | 79.7 | 59.4 | 61.1 | 68.4 | 79.3 | 73.3 |
| MCNet | 91.8 | 81.2 | 96.9 | 98.4 | 93.7 | 91.8 | 92.8 |

**Table 3. The Result on the DSSE-200 Dataset. The DV3+ Means the DeeplabV3+ Model with Xception Backbone**

| Method | A | P | R | F1 |
|---|---|---|---|---|
| FCN [10] | 0.69 | 0.68 | 0.64 | 0.66 |
| Segnet [11] | 0.76 | 0.71 | 0.72 | 0.71 |
| PANet [12] | 0.79 | 0.74 | 0.72 | 0.73 |
| PSPnet [13] | 0.72 | 0.69 | 0.79 | 0.74 |
| DV3+ [14] | 0.78 | 0.72 | 0.75 | 0.73 |
| MCNet | 0.92 | 0.88 | 0.85 | 0.81 |

**Table 4. Ablation Experiments on DSSE-200.**

| Method | Acc | P | R | F1 |
|---|---|---|---|---|
| MCNet | 0.925 | 0.883 | 0.856 | 0.812 |
| Two-stream | 0.921 | 0.893 | 0.842 | 0.810 |
| 6-Channel | 0.924 | 0.882 | 0.849 | 0.804 |

To begin with, we evaluated our model’s architecture by comparing it with the prevalent dual-stream models in the current research landscape. Unlike the dual-stream approach, which refrains from directly combining the Mask and RGB channels and instead opts for extracting features from the Mask map using a similar structure before proceeding with the fusion, our model adopts a strategy of direct fusion. The comparative analysis, as detailed in Table 4, reveals that our model’s direct fusion methodology outperforms the dual-stream structure. This superiority is attributed to the preservation of crucial information that is otherwise lost during the layer superposition process in the dual-stream model, which, in turn, affects the formation of positive feedback.

Furthermore, the decision to limit the input to four channels in our Mask Constraint module is also derived from empirical evidence. As indicated in Table 4, the performance achieved with six channels was found to be comparable to that with four channels. However, the use of six channels resulted in an increase in the model’ sparameters without yielding proportional benefits in terms of effectiveness. Based on these findings, we opted for a four-channel input to optimize the balance between model complexity and performance efficiency. This choice underscores our commitment to developing a model that not only delivers superior performance but also maintains a lean architecture, thus facilitating easier adaptation and application in diverse document layout analysis tasks.

**MASK Embedding Block.** As previously discussed, we introduced the Mask Constraint (MC) module specifically to tackle the semantic gap issue frequently encountered in the application of Mask R-CNN to document layout analysis. To illustrate the effectiveness of this module, we conducted a comparative analysis involving samples processed by both Mask R-CNN and our MCNet. The comparison, depicted in Figure 5, clearly demonstrates the capability of MCNet to address the problem of object discontinuity that arises due to semantic gaps. The samples processed by MCNet show a significant improvement in maintaining the continuity of document objects, effectively bridging the semantic gap that poses challenges for conventional Mask R-CNN applications. This evidence underscores the value of the MC module in enhancing the performance of document layout analysis by ensuring a more coherent and accurate segmentation of document elements.

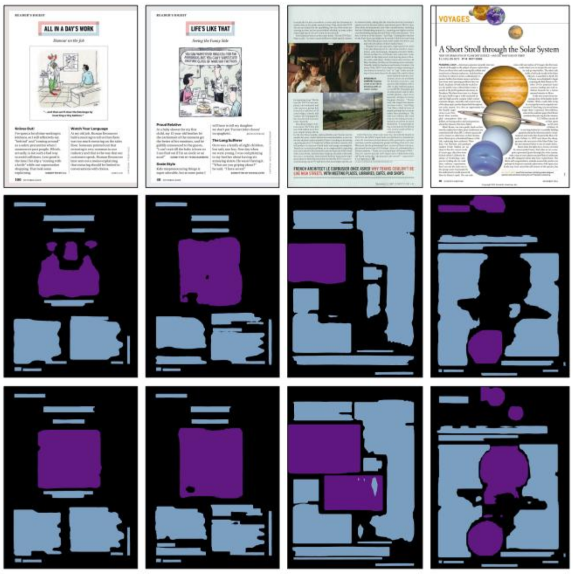

**Figure 5. Example Real Documents and Their Corresponding Segmentation on FPD. Top: Original. Middle: Mask R-CNN. Bottom: MCNet.**

## 4. Conclusions

In this paper, we propose a Mask-constrained document layout analysis framework to deal with the semantic gap defect of document layout analysis processing using detection methods. The proposed Mask constraint module is a lightweight component. In addition, we propose a Constrained Aggregation algorithm for the problem of overlapping regions in detection methods. We demonstrate the saliency of the proposed framework on DSSE-200 and FPD datasets.

## References

[1] Yang, X.; Yumer, E.; Asente, P.; Kraley, M.; Kifer, D.; Lee Giles, C. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5315–5324.

[2] Clark, C.; Divvala, S. Pdffigures 2.0: Mining figures from research papers. In Proceedings of the ACM/IEEE on Joint Conference on Digital Libraries, 2016, pp. 143–152.

[3] Clark, C.A.; Divvala, S. Looking beyond text: Extracting figures, tables and captions from computer science papers. In Proceedings of the Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[4] Praczyk, P.A.; Nogueras-Iso,J. Automatic extraction of figures from scientific publications in high-energy physics. Information Technology and Libraries 2013, 32, 25–52.

[5] Li, Y.; Zou, Y.; Ma, J. Deeplayout: A semantic segmentation approach to page layout analysis. In Proceedings of the International Conference on Intelligent Computing. Springer, 2018, pp. 266–277.

[6] Ma, W.; Zhang, H.; Jin, L.; Wu, S.; Wang, J.; Wang, Y. Joint layout analysis, character detection and recognition for historical document digitization. In Proceedings of the International Conference on Frontiers in Handwriting Recognition (ICFHR). IEEE, 2020, pp. 31–36.

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

[8] He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence 2015, 37, 1904–1916.

[9] Ma, T.; Wu, X.; Du, X.; Wang, Y.; Jin, C. Image Layer Modeling for Complex Document Layout Generation. In Proceedings of the 2023 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2023, pp. 2261–2266.

[10] Loc, C.V.; Burie, J.C.; Ogier, J.M. Document images watermarking for security issue using fully convolutional networks. In Proceedings of the ICPR, 2018, pp. 1091–1096.

[11] Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 2017, 39, 2481–2495.

[12] Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid attention network for semantic segmentation. British Machine Vision Conference 2018.

[13] Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia,J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.

[14] Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision, 2018, pp. 801–818.