Noname manuscript No.
(will be inserted by the editor)

Beyond Document Object Detection: Instance-Level Segmentation of Complex Layouts

Sanket Biswas · Pau Riba · Josep Lladós · Umapada Pal

Received: date / Accepted: date

Abstract Information extraction is a fundamental task of many business intelligence services that entail massive document processing. Understanding a document page structure in terms of its layout provides contextual support which is helpful in the semantic interpretation of the document terms. In this paper, inspired by the progress of deep learning methodologies applied to the task of object recognition, we transfer these models to the specific case of document object detection, reformulating the traditional problem of document layout analysis. Moreover, we importantly contribute to prior arts by defining the task of instance segmentation on the document image domain. An instance segmentation paradigm is especially important in complex layouts whose contents should interact for the proper rendering of the page, i.e., the proper text wrapping around an image. Finally, we provide an extensive evaluation, both qualitative and quantitative, that demonstrates the superior performance of the proposed methodology over the current state-of-the-art.

Keywords Document Object Detection · Layout Analysis · Instance Segmentation · Information Extraction

S. Biswas, P. Riba, J. Lladós
Computer Vision Center & Computer Science Department
Universitat Autònoma de Barcelona
Tel.: +34-935811828
E-mail: {sbiswas, priba, josep}@cvc.uab.es

U. Pal
CVPR Unit, Indian Statistical Institute, India
E-mail: umapada@isical.ac.in

1 Introduction

Visual perception is one of the most fundamental senses for humans and others animals. It is the power of vision that enables us (humans) to interact with the environment. To actually reproduce the human visual perception abilities in Intelligent Systems the design of computational visual models is required. A small step towards this goal in computer vision is by training a neural network model to learn the essential image components (features) that are significant and interesting to human observers to search for a definite object category. The task of object detection in Computer Vision can be analogously stated in the domain of Document Image Analysis and Recognition (DIAR). Thus, Document Object Detection (DOD) is the task of decomposing the image into semantically meaningful regions such as tables, figures, titles, paragraphs, etc. [25]. Classically, this concept was referred as logical layout segmentation. Object detection in Computer Vision has grown immensely in recent years with end-to-end trainable models deployed on deep CNNs. Equivalently, in DIAR, tasks such as preprocessing [18, 22], layout analysis [47, 49], character segmentation [41, 57], and signature verification [8, 19] have benefited so much from extremely robust state-of-the-art machine learning techniques practised in recent years.

Twenty years ago, the documents of the future were seen as digitally born, accessible, indexable, etc. New paradigms of information management workflows in organizations, without the need of scanning documents were predicted. This has been partially fulfilled. Sellen and Harper described this prediction as the Myth of the Paperless Office [43] and stated that paper-based documents would continue to play an important role in office life. Nowadays, the processing of (image) docu-

ments has not only grown, but the advent of new digital services in different sectors (e.g. fintech, legaltech, insurtech) has resulted in new challenges for the document interpretation task, combining both sources, paper and digitally born. Thus, with the rapid increase in the usage of digitized documents over the years, the need for automated methods for extraction and retrieval of information has become a necessity. Manual approaches are no longer a feasible option. There has been tools and applications in recent times to convert these digital documents into process-able entities. It has been observed that semantically segmenting page elements such as tables, figures, paragraphs, and mathematical equations can indeed play a vital role in understanding and extracting information from documents. Document Object Detection (DOD) aims to automatically model a document page into its structural and logical graphic entities for its application to solve a number of document image analysis tasks. These specific tasks range from document content understanding, document structural and syntactic analysis [6, 42, 48] and so on. A document structure cannot be explicitly encoded by the two most popular document formats, images and PDFs. While images encode pixels, PDFs encode vector, raster and text marker information. Therefore, detection of graphical elements and structural objects in digitally generated documents is a challenging and interesting problem. Layouts play a significant role in dictating the reader's attention and hence, the order by which it conveys the information. Variations in document layout can often change the hierarchy and narrative of the information. In this work, we have therefore focused on detecting and recognizing graphical objects and deducing their spatial and structural relationships to solve the problem of understanding document layouts for information extraction.

Recent advances in object detection for natural scene images [28, 37] has captured a lot of attention. Our problem to detect objects and understand layouts in documents is conceptually quite similar. But the large domain discrepancies in document images make it quite challenging to apply in this scenario compared to natural scene images. The diversity in aspect ratio and scale of document objects and named entities are far more significant as compared to that in natural scene objects. For example, in case of scientific articles, tables may occupy the most part of a page, logos or figures may appear on any side of the page column width while lines of text or paragraphs can have an extreme aspect ratio. This inter-class variance in the structure of graphical layouts in documents makes its detection quite difficult. Rule-based systems [7, 9, 44, 52] used before the advent of deep learning considerably fail to detect these document objects in such variable test case scenarios. In recent years, deep Convolutional Neural Networks (CNNs) have been able to create a major improvement in detection performance for such diverse graphical objects [24, 36]. However, most of these deep CNN-based methods generally try to deduct the visual differences between the object classes: while the visual characteristics of certain graphical elements (e.g. plots, charts) differ conspicuously from text, the same cannot be said for tables, where the major differences from the surrounding content lie mostly stored in the layout information and its context.

In general, object detection provides the classes of the objects and their location based on bounding boxes. *Instance segmentation* is a more precise task that provides the boundaries of the objects at the detailed pixel level. The difference is important when there are complex layouts where text, figures or other objects overlap. Figure 1 displays an example image of a historical document and a scientific article where there is a gross overlapping between object categories, and extracting this layout structure is difficult with only bounding box information. On the left historical document image in the figure, the title and subtitle information appears inside a row, while on the right the text blocks appear inside the list category in scientific documents. These case studies raise the requirement to solve the problem of extracting layout information with instance masks added to the bounding box information.

In this work, we have therefore focused on going beyond object detection for understanding document layouts. The basic idea is to add another segmentation module to state-of-the-art document object detection systems that is able to generate segmentation masks for every individual object category of a document image. The motivation behind this idea was the release of some very significant large scale annotated datasets [45, 61] to the document analysis community in recent times by prominent research groups. Surprisingly, no steps have been taken to use this mask-level information for parsing the spatial layout information in digitized documents. For example, historical documents in Asian languages have really unique and complex structured layout patterns which are hard to comprehend. In such cases, just using defined bounding box information of layout objects may not be useful. Instead, we require a more robust and scalable system that can isolate the individual instances of each region (eg. title region, text lines, etc.) in the document image, and prevent overlapping of regions that represent a hierarchical structure as shown in Figure 2. A similar scenario can also appear in the pages of scientific articles, where a table may appear inside the figure region. Therefore, we

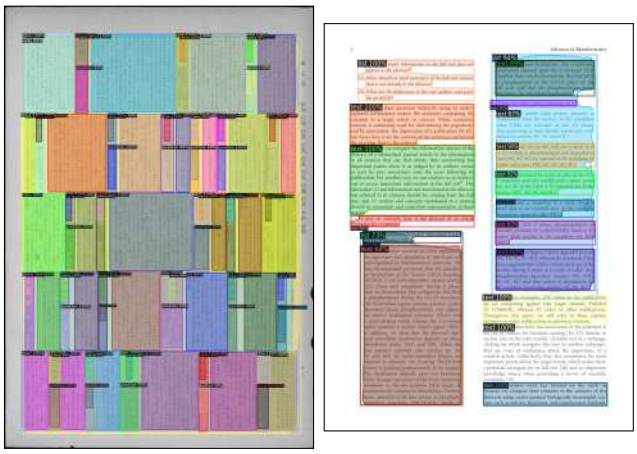


Fig. 1: A sample illustration from the HJDataset (left) and PubLayNet (right) with overlapping object categories.

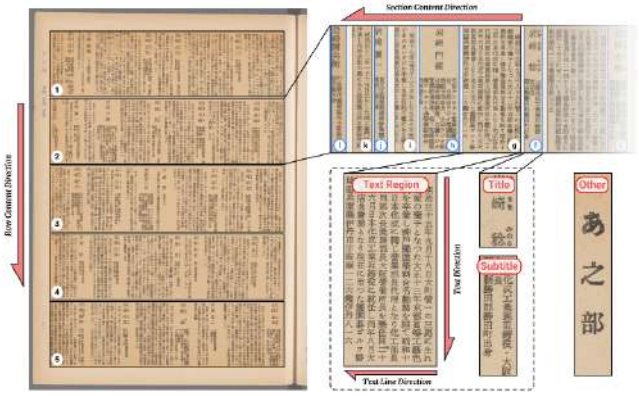


Fig. 2: Hierarchical content structure as illustrated in the Historical Japanese document dataset by Shen et. al. [45].

formulate the understanding of spatial layouts in structured documents as an instance segmentation problem.

The contributions of this work can be divided into three folds:

- We establish important baselines using our proposed instance segmentation model based on the Mask-RCNN [16, 17] architecture on two recent benchmark document datasets, the PubLayNet [61] and the Historical Japanese Dataset [45]. To the best of our knowledge, this is the first work addressing this benchmark task to understand and analyze complex document layouts. We have tried to motivate the idea of going from bounding box-level to instance-level segmentation of complex document layouts in document object detection.
- We also adapt and evaluate our instance-level segmentation framework to document object detection tasks on these datasets and compare with the existing state-of-the-art approaches. The conducted experiments in this study prove that our model effectively advances the current state-of-the-art.
- We also propose several interesting ablation studies to justify the effectiveness and impact of our instance segmentation model on both document object detection and instance segmentation tasks.

The rest of the paper is organized as follows. In Section 2 we review the state-of-art methods related to object detection and instance segmentation, specially focused on the document image analysis domain. Section 3 explains in detail our proposed method for document understanding at segmentation and detection levels. Afterwards, Section 4 performs an extensive evaluation for both tasks. In Section 5 we discuss the achieved results. Finally, Section 6 draws the conclusions and proposes open challenges for further research.

2 Related Work

Automatic information extraction from digital documents requires the understanding of their spatial layout elements. This involves the detection of semantically meaningful objects such as tables, titles, figures, text blocks, etc. Several works have approached the localization of page objects and the analysis of the page spatial layout. In this section, we overview the state-of-the-art according to different methodological schemes.

2.1 *Traditional Document Layout Segmentation*

Identifying the structure of digital documents is a well-known research problem. There has been several hand-crafted rule-based segmentation approaches that have been carried out in the past. Binmakhashen *et. al.* [5] provided a vigorous survey on several approaches for the extraction of physical layout and detecting logical structures in document images. According to them, these methods in the literature can be categorised as top-down, bottom-up and hybrid approaches.

In general, bottom-up strategy initiates on a lower level of an image such as pixels, words, or components, and then it evolves to a higher level structure like document regions and stops once it reaches a predefined analysis objective. O'Gorman [32] studied this problem by grouping connected components on polar structural parameters (angles, distances etc.) to derive the segmentation. His work became quite successful and popular as the Docstrum algorithm. Kise *et. al.* [21]

used the area Voronoi diagram for segmenting page images. However, the Voronoi algorithm is computationally quite expensive when applied to these digital documents. To address this problem, Delaunay triangulation was employed to solve text line segmentation [21] and extracting author and title regions [55]. Journet *et. al.* [20] further used a spatial autocorrelation approach to highlight some periodicities and texture orientation for segmenting graphic elements in a page. This autocorrelation approach performs better for a document having a complex layout or text written in various fonts. Agrawal *et. al.* [1] proposed an upgrade to bottom-up layout analysis by integrating Docstrum algorithm with Voronoi algorithm to track neighboring components and producing a better regional segmentation of spatial layouts.

There has also been some relevant work on segmenting spatial layouts with well defined top-down strategies. Ramel *et. al.* [35] used a white-space analysis approach to detect regions that can be segmented by spaces (i.e. background) from all directions. Saabni and El-Sana [40] proposed a method that developed seem lines among text lines using energy maps. Asi *et. al.* [2] came up with a multi-scale texture based algorithm for document images where Gabor filters were applied to locate different regions and a minimization energy function was applied to segment them. Despite successes of both top-down and bottom-up strategies, there are techniques [51] that have integrated both of them to segment regions in digital documents with complex layouts.

Prior to the deep learning era, most of the rule-based segmentation approaches [7, 9, 44, 52] also aimed to solve the table detection problem by assuming table structures and using the prior knowledge on the object attributes and extracting them by analyzing tokens from documents. Apart from rule-based approaches, several studies [4, 31, 53] were also conducted to consider machine learning for document understanding which included pre-processing, segmentation and labeling. The most notable work in this case is the multilayer perceptron (MLP) network used by Marinai et. al. [31]. But the key disadvantage of all these approaches was they failed to generate good results in digital documents with variable layout structures. The structure and type of the particular layout objects (tables, text lines, etc.) were taken as inherent assumptions for using these segmentation based approaches. This incensed the need for data driven approaches using deep CNN's in the document analysis community for providing a more robust solution to solve its tasks.

2.2 *Document Object Detection with Deep Learning approaches*

With the advent of deep CNN's, the performance of object detection tasks in computer vision has achieved manifolds. It is always recommended to have a strong backbone feature extractor for building accurate detection models. In recent state-of-the-art, custom object detection networks are broadly classified into two different categories: two-stage and one-stage. Faster-RCNN (FRCNN) [37] is one of the most popular two-stage detectors used in recent years for object detection tasks. It generates coarse-grained object proposals using a Region Proposal Network (RPN) module in the first stage. These region proposals and refined features are then fed to the classification module in the second and final stage. In this work, FRCNN has been used as a baseline for comparison with our proposed model. There also exist other state-of-the-art two-stage [12] and one-stage object detectors like SSD [30], YOLO [36] and Retinanet [29].

In recent years, deep CNNs have proved to be quite effective and have been explored by quite a number of research groups in document analysis. Hao *et. al.* [13] have used them for detecting tables in PDF documents by generating region proposals with table-like structures in document pages and then classifying them into table or non-table entities using a CNN. Augusto *et. al.* [3] proposed a fast one-dimensional document layout analysis approach based on CNNs to segment text, figures and tables in a page. Schreiber *et. al.* [42] then devised a very standard deep learning approach called DeepDeSRT for table detection and structure recognition, where no prior knowledge or assumption about table structures was necessary. In contrast to the existing table detection and structure recognition approaches that were only applicable for processing PDFs, DeepDeSRT helped processing document images too which made it quite robust to process born-digital images as well as even harder problems, e.g. scanned documents. They applied fine-tuning on existing Faster-RCNN with emphasis on two different backbones for table detection task: ZFNet [60] and VGG16 [46]. Gilani [11] improved the model performance further for table detection without changing the model backbones used by Schreiber *et. al.* [42]. They introduced a new pre-processing unit where they applied image transformation on the samples using a stack of 3 different distance transformed layers before feeding them to the Faster-RCNN. He *et. al.* [14] used a two-stage system architecture for detecting table and figures. The class label for every pixel is predicted using a multi-scale and multi-task Fully Convolutional Neural Networks (FCNN) in the first stage

of the model framework. The next stage is the region proposal network where certain heuristic rules are applied on these pixel-wise class predictions to get the object bounding boxes (region proposals). Oliviera *et. al.* [33] adopted a similar FCNN-based approach for pixel-wise segnmentation on historical documents. Gao *et. al.* devised an end-to-end approach to detect mathematical formulas using a combination of CNNs and RNNs (Recurrent Neural Networks) to use both character and vision features. They extracted the meta-data information from the PDF files before passing them to the feature extraction network. To extend the problem to detect multiple graphical objects jointly, Gao *et. al.* [10] organized a competition in the 2017 edition of the International Conference in Document Analysis and Recognition (ICDAR) and proposed a baseline for multi-object scenario. Yi *et. al.* [59] redesigned the common CNN object detection approach with a newly devised training strategy, network structure and used a dynamic programming algorithm to prevent the usage of Non Maximal Suppression (NMS) for their model training. Li *et. al.* [26] used a deep structure prediction to obtain primitive region proposals from every column region. Then these primitive proposals are clustered and merged with Conditional Random Field (CRF) based graphical models that can be used to integrate both contextual and local information. Riba *et. al.* [38] used the geometric relations between layout elements to parse tables in administrative invoice documents. Xu *et. al.* [58] designed a framework called LayoutLM to model joint interactions between layout and text to understand structured documents like administrative forms and receipts.

In recent times, Zhong *et. al.* [61] introduced a novel large scaled annotated dataset for document layout analysis tasks in scientific literature called PubLayNet and also introduced some state-of-the-art baselines. It includes both bounding boxes and mask information for regional graphic objects in a page. Shen *et. al.* [45] introduced another large scaled labeled dataset on historical Japanese documents called HJDataset that also contained bounding box and mask information of the layout elements and also included a hierarchical structure and reading orders for those elements. Both these datasets in the literature helped to evolve real-world document digitization tasks on both scientific and historical literature. In this work, we decided to focus on the usage of the mask information of layout elements to a new instance-level segmentation task for documents. The most popular and efficient state-of-the-art instance segmentation model is the Mask-RCNN [16]. Huang *et. al.* [17] provided an improved mask scoring strategy to encourage more accurate mask predictions for the instance segmentation task. Inspired by both of these works, we propose a simple and effective approach that serves as a solid baseline and ease future research in instance-level segmentation task for document images. We have also compared this model with existing state-of-the-art document object detection baselines for the corresponding datasets, PubLayNet [61] and HJDataset [45].

3 Instance-level segmentation model

This section presents a complete analysis of our proposed end-to-end instance-level segmentation model that has been inspired from state-of-the-art instance segmentation models, Mask-RCNN [16] and Mask Scoring RCNN [17]. Detecting layout objects in documents that include tables, figures, paragraphs, title etc. is a well-studied problem. Powerful CNN architectures [29, 37] have demonstrated remarkable performances in detecting these document layout elements accurately. In order to move from bounding box-level object detection to a more accurate pixel-level classification, a novel framework has been introduced to utilize and evaluate instance-level segmentation of layout elements in documents. Figure 3 shows a detailed overview of the proposed framework for instance-level segmentation of document layout objects. This framework has been explained vividly in four different modules: (i) feature extraction and selection module; (ii) object detection head; (iii) instance segmentation head; and (iv) learning objectives.

3.1 Feature extraction and selection module

Similar to previous approaches for object detection and segmentation, the proposed model also adapts a convolutional backbone for extracting image features. We have adapted the ResNeXt-101 [56] aggregated backbone for our model. The ResNeXt-101 has an internal dimension for each convolutional path denoted as d ($d = 8$). The number of paths is represented as the cardinality C ($C = 32$). As we aggregate the dimension of each 3×3 convolution (*i.e.* $d \times C = 8 \times 32$), it gives us 256 features. So we provide an input image of a document into the ResNeXt-101 base CNN to get a feature hierarchy from the convolutional layers. The output of the convolutional stack is then input into a Feature Pyramid Network (FPN) which exploits the multi-scale feature representation obtained by the CNN block to build a richer semantic representation. FPN iterates from the most coarse feature map, upsamples it by a scale factor of 2 for enhancing spatial

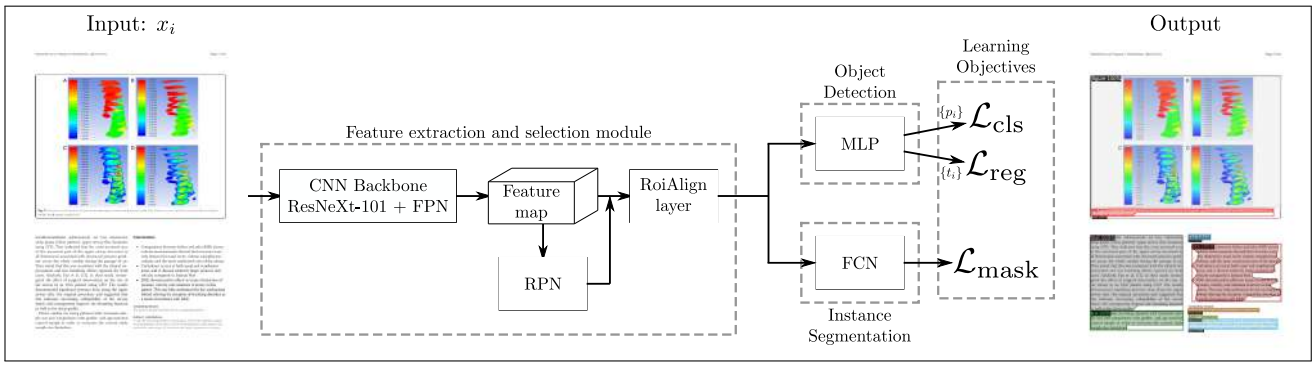


Fig. 3: **Proposed Instance-Level Segmentation framework:** Given an input image of a document, the model predicts the different layout elements, with object detection on one head and instance-level segmentation on another head.

resolution, and then merges it with the preceding map, that has already undergone a 1×1 convolution. The merged feature is then smoothed with 3×3 convolution to obtain the final feature map. The final feature map has a dimension of 512.

The feature pyramids help to extract multi-scale feature maps with different receptive fields which not only combine high spatial resolution information from early in the stack with low-resolution but also rich semantic information that is required from deeper in the stack. This rich semantic information is especially really helpful for detecting wide categories of objects in document images, both small as well as big objects.

The Region Proposal Networks (RPN) module proposes rectangular regions, each associated with objectness score that tells whether it represents an object or not. The region proposals are extracted from all feature pyramid layers by the RPN and chooses the top 1,000 precomputed 'region proposal' boxes using the predicted objectness scores at each feature level for an image and ranking them after using Non-Maximal Suppression (NMS) independently on top of it. The proposals higher than the Intersection over Union(IoU) threshold applied during training are counted as foreground (object) and the rest as background.

We take the feature maps from FPN, the top 1,000 proposal boxes from the RPN and the ground truth input boxes and feed them to the Region of Interest (RoI) Align layer. The RoIAlign layer tries to align properly the extracted features with the input. A bilinear interpolation strategy is used to calculate the appropriate values of the input features at each RoI neighbourhood. This layer provides an advantage over the RoI Pooling layer applied in the Faster-RCNN detector [37] as they avoid quantization operations (eg. floor, ceil) to map the generated region proposals with exact x-y integer indexes. These operations tend to generate a misalignment between the extracted features and the RoI which might lead to a negative impact on predicting pixel-accurate masks during segmentation stage. Given the bilinear interpolated points in feature layers and a list of proposals which are scaled by a scaling factor, the RoI Align layer returns an aligned representation of the proposals. The obtained results are then concatenated to generate a fixed size feature map by preserving the spatial locations. This feature map is then fed to the two model heads: object detection head and Instance segmentation head, as shown in Figure 3.

3.2 Object detection head

Given the aligned feature map the object detection branch aims at detecting and recognizing the different blocks that compose our page layout. In particular, the object detection head, defined as a Multilayer Perceptron (MLP), predicts per each anchor, two outputs. On the one hand, a classification score in terms of the different object categories like tables, figures, title, text blocks, etc. and, on the other hand the 4 coordinates of the corresponding bounding box of those categories.

This module is defined as a two fully-connected layers with ReLU activation function.

3.3 Instance segmentation head

A Fully Convolutional Network (FCN) has been proposed as the instance segmentation head. This network predicts the final binary mask for each one of the possible object categories in the selected RoIs. It has been observed that the instance segmentation head performs a pixel-based classification.

3.4 Learning objectives

Three learning objectives are proposed to guide the learning process of the proposed network, two of them associated to the object detection head, and the third one related to the segmentation task. Thus the learning objective is formally defined as

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{mask}}, \tag{1}$$

where $\mathcal{L}_{\text{det}}$ and $\mathcal{L}_{\text{mask}}$ are the detection and segmentation losses respectively.

3.4.1 Detection objective

The first component of our learning objective is the detection loss $\mathcal{L}_{\text{det}}$. It follows the same strategy as the one proposed for the Faster-RCNN [37] object detection architecture. Given an image, the proposed model predicts a set of probabilities $\{p_i\}_{i=0}^{k}$ which is the predicted probability of anchor i being or not an object, and a set of 4 parameterized coordinates $\{t_i\}_{i=0}^{k}$ of the predicted bounding box. Following the traditional formulation of supervised object detection, we know beforehand the corresponding ground-truth for each of the sets, namely, $\{p_i^*\}_{i=0}^{k}$ and $\{t_i^*\}_{i=0}^{k}$. Note that p_i^* is 1 if the corresponding anchor is an object, and is 0 otherwise. The detection loss is formally define as

$$\begin{aligned}\mathcal{L}_{\text{det}}\left(\{p_i\},\{t_i\}\right) = &\frac{1}{N_{\text{cls}}}\sum_i \mathcal{L}_{\text{cls}}\left(p_i, p_i^*\right) \\ &+ \frac{\lambda}{N_{\text{reg}}}\sum_i p_i^* \cdot \mathcal{L}_{\text{reg}}\left(t_i, t_i^*\right),\end{aligned} \tag{2}$$

where $\mathcal{L}_{\text{cls}}$ stands for the object classification loss and $\mathcal{L}_{\text{reg}}$ the bounding box regression objective. These losses are defined as the binary cross entropy and the smooth L_1 loss.

That is,

$$\mathcal{L}_{\text{cls}}(p, p^*) = \left[p^* \log p + (1-p^*)\log(1-p)\right], \tag{3}$$

$$\mathcal{L}_{\text{reg}}(t, t^*) = \begin{cases} 0.5(t-t^*)^2/\beta & \text{if } |p-p^*|, < \beta \\ |t-t^*| - 0.5\beta, & \text{otherwise} \end{cases} \tag{4}$$

where β is a threshold parameter set to 1, $p \in \{p_i\}$, $t \in \{t_i\}$ and, p^* and t^* are their corresponding groundtruths.

3.4.2 Segmentation objective

In addition to the detection loss, a segmentation objective is incorporated in order to obtain a fine-grained mask of our document instances. Therefore, the mask loss proposed in the Mask RCNN architecture [16] is used. Note that this loss is computed per each one of our object categories.

$$\begin{aligned}\mathcal{L}_{\text{mask}}^{k}(y) = -\frac{1}{m^2}\sum_{1\le i,j\le m} &\left[y_{ij}^* \log y_{ij}^k \right. \\ &\left. + (1-y_{ij}^*)\log(1-y_{ij}^k)\right],\end{aligned} \tag{5}$$

where y_{ij}^* is the ground-truth of the category k of the cell (i,j) in the mask for the region of size $m \times m$ and y_{ij}^k is the predicted value of the same cell and class k.

4 Experimental Validation

For validating purposes, we have considered significant benchmark datasets with different document typologies. Our empirical evaluation demonstrates that the proposed approach provides competitive results when compared to the state-of-the-art. Moreover, extensive ablation studies show the contribution of each one of the modules. All code and proposed benchmark models will be publicly available at: `https://github.com/biswassanket/instasegdoc`.

4.1 Evaluation Measures

A common way to determine the correctness of an object proposal is the Intersection over Union (IoU). In this work, we use the mean average precision (mAP) metric calculated by averaging the average precision (AP) at different IoU thresholds ranging from 0.5 to 0.95 with a step size of 0.05. This is the primary metric used in the standard MS-COCO [27] dataset for evaluating the performances of state-of-the-art object detection and instance segmentation models. Moreover, we provide AP scores over IoU thresholds 0.5 (AP@0.5) and 0.75 (AP@0.75) for evaluating the tasks in our proposed model. All model performances have been evaluated both categorically and with an overall average AP score based on AP scores for each category or object.

4.2 Dataset Description

During the past years, there has always been a lack of publicly available datasets that concerns the structural extraction of complex layouts from modern digital documents. The main reason lies in the sensitiv-

ity of its contents which make them strictly confidential. Thus, the efforts of several research groups allowed the research community to contribute such significant datasets in recent times. In this work, two newly released datasets have been considered for the evaluation of our approach namely, PubLayNet [61] and HJ Dataset [45].

4.2.1 PubLayNet

The PubLayNet dataset [61] for Document Layout Analysis task was launched in the International Conference for Document Analysis and Recognition (ICDAR2019) and provided one of the breakthrough contributions to the Document Analysis community. The size of this dataset is also comparable to large-scale computer vision datasets, with over 360 thousand images taken from the list of PDF articles present in the PubMed Central [39] library for scientific literature. The defined sets of categories to detect in this dataset are text, title, lists, tables, and figures. The dataset has been used for both training and evaluation purposes in our study and a complete summary of its object categories is shown in Table 1. It is also provided with ground truth masks that actually help us to evaluate instance segmentation performance of our model. The entire dataset has been trained for 335,703 images and 11,245 images for evaluation. We used the PubLayNet validation set for evaluation as we could not use the official testing set as the ground-truth has not been released due to an ongoing competition.

Table 1: Summary of the PubLayNet dataset used for our experimental evaluation.

Object Category	# Instances	
	Train	Evaluation
Text	2,343,356	88,625
Title	627,125	18,801
Lists	80,759	4,239
Figures	109,292	4,327
Tables	102,514	4,769
Total samples	3,263,046	120,761

4.2.2 HJDataset

The Historical Japanese documents (HJDataset) dataset [45] contains 2,048 images with 250,000 layout element annotations of seven different categories: from page frames to individual text blocks. It also provides a relevant ground truth information for these objects in the form of bounding boxes and masks that has been used for our model evaluation study. It contains various document information of 50,000 prominent Japanese citizens from the Japanese Who's Who biographical directory. In our study, the data has been distributed into 1,433 images for training, 307 images for validation and 308 images for final testing. A summary of the distribution of categorical instances used during training and evaluation are shown in Table 2.

Table 2: Summary of the HJDataset used for our experimental evaluation.

Object Category	# Instances	
	Train	Evaluation
Body	1,443	308
Row	7,742	1,538
Title	33,637	7,271
Bio	38,034	8,207
Name	66,515	7,257
Position	33,576	7,256
Other	103	29
Total samples	181,097	31,866

4.3 Ablation Study

Extensive ablation studies were conducted in the context of document object detection to quantify the significance of every component of our overall model framework and justify its usage for both detecting and segmenting different layout elements.

4.3.1 ResNet vs ResNeXt

Having originated from ResNet [15], VGG [46] and Inception [50], ResNeXt-101 [56] models gave a better performance in object detection as compared to ResNet-101 [50]. A study was conducted with the baseline Faster-RCNN and Mask-RCNN models evaluated on the PubLayNet dataset and the results on the overall mAP as shown in Table 3 clearly justify the usage of the ResNeXt-101 backbone for the final model.

4.3.2 Use of FPNs

To evaluate the impact and effectiveness of Feature Pyramid Networks (FPNs), we conducted experiments with the Faster-RCNN and Mask-RCNN models with the default ResNeXt-101 backbone on the PubLayNet dataset. As shown in Table 4, FPN actually helps to

Table 3: Study on the convolutional backbone for our detection and segmentation framework in terms of mean Average Precision.

Model	Backbone	
	ResNet-101 [50]	ResNeXt-101 [56]
F-RCNN	0.828	0.843
M-RCNN	0.869	0.875

provide a substantial gain in overall mAP score for object detection for both the baseline models consistently, which highlights the importance of usng FPN's in our model backbone.

Table 4: Performance analysis on the PubLayNet dataset of the DOD model with or without FPN.

Model	mAP
F-RCNN	0.843
F-RCNN (+ FPN)	0.871
M-RCNN	0.875
M-RCNN (+ FPN)	0.904

4.4 Implementation details

The model has been implemented using the popular detectron2 [54] framework released by the Facebook research team for end-to-end state-of-the-art object detector models. We use this framework as it is built on top of PyTorch [34] and is extremely simple and robust for model deployment. Nvidia Titan X GPU's have been utilized for all our training purposes. Pre-trained model weights of ResNeXt-101 [56], trained on ImageNet [23] dataset, have been used as backbones for our trained models.

To fine tune, we initialize the weights on the pre-trained models and we train the "heads" layer using our dataset. The model ran upto a total of 30,000 iterations with an initial learning rate of 0.00025. To generate k=32 anchor boxes, we considered different anchor scales to cover almost all parts of the image. Stochastic Gradient Descent (SGD) optimizer was used to train with Nesterov Momentum was used with a batch size per image of 128 in the RoI heads. The learning rate scheduled with Warmup Cosine Annealing after every 10,000 iterations of training. The detection minimum confidence score was set to 0.7 for our task and the detection NMS threshold was set to 0.3. The number of workers in the dataloader was set to 4. After the fine-tuning is completed, we set the testing threshold in the RoI heads for the model to be 0.6. We tried different testing thresholds and achieved our best results with this score. Moreover during the training time, we used the default data augmentation strategy in the detectron 2 framework that uses random flipping both vertically and horizontally.

Table 5: Results for the PubLayNet dataset for the tasks of Document Object Detection and Document Instance Segmentation.

Category	Detection			Segmentation
	F-RCNN	M-RCNN	Ours	Ours
Text	0.910	0.916	0.918	0.906
Title	0.826	0.840	0.844	0.818
List	0.883	0.886	0.913	0.821
Table	0.954	0.960	0.971	0.970
Figure	0.937	0.949	0.951	0.948
AP	0.902	0.910	0.920	0.893
AP@0.5	-	-	0.977	0.977
AP@0.75	-	-	0.959	0.953

5 Results and Discussions

In this section, we have tried to provide some deep insights into the experimental results we have achieved both qualitatively and quantitatively. We shall discuss them in the following lines according to the datasets we have used to evaluate our proposed method.

5.1 Results on PubLayNet

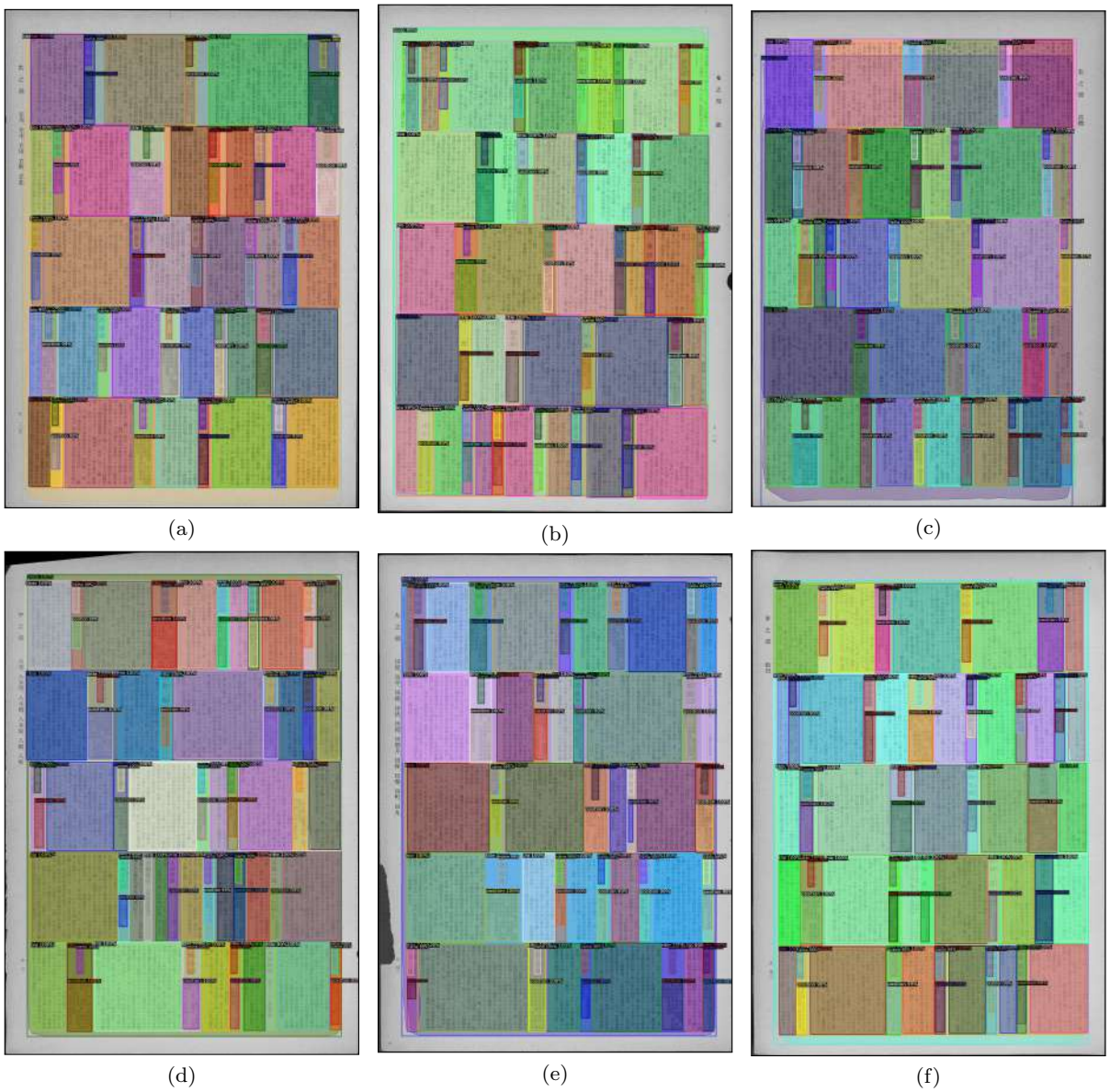
The results after training and evaluating our proposed instance-level segmentation model on the PubLayNet dataset are shown in Table 5. The mAP score has been computed in this case for all different categories of objects (text, lists, tables, title, and figures) that have been detected and segmented as shown in the result table. We have also provided a new instance segmentation baseline on the evaluation of predicted masks by our proposed model, which is a novel contribution. In addition, the results for object detection obtained by our model have been compared with previous baselines provided by state-of-the-art models. We achieve the state-of-the-art results on detection performance with an overall AP of 0.92. Furthermore, the AP scores obtained for individual object categories are relatively higher compared to the existing baselines on Faster-RCNN and Mask-RCNN proposed by Zhong *et. al.* [61]. While on our instance segmentation baseline, our model has an overall AP score of 0.893.

(a) (b) (c)

(d) (e) (f)

Fig. 4: Qualitative analysis on the PubLayNet dataset by the instance-level segmentation model.

Very high AP scores for object detection has been obtained for the object categories table, figure, list and text blocks, which is really encouraging. The relatively low AP for title detection may be attributed to the fact that our model can still improve the performance for detecting smaller objects. Also, the title category seems to be the weakest due to the variability of its representation in different pages of the scientific articles. On the other hand, titles often get misclassified into a text block category when there is not much spacing between the two elements in the document layout structure. As expected, the AP scores for instance segmentation on the title is relatively lower with 0.81. However, the relatively low AP score of the list category for instance segmentation makes sense due to more false positives arising between list objects and text blocks. Figure 4(c) and Figure 4(d) clearly illustrate this problem for list and text blocks. Figures 4(b) and 4(c) shows perfect detection and segmentation of overlapped object categories. In Figure 4(b), tables inside the figure region gets correctly segmented and detected. There are two overlapped objects in this figure and both of them gets correctly segmented with a perfect AP score. While in Figure 4(c) small object categories like the captions of figures get correctly segmented and detected although they lie inside the figure region. In this case too we observe a perfect segmentation score for overlapped objects. Overall, the qualitative results shown in Figure 4 clearly illustrate how our model performs in variable cases of layout organization.



(a) (b) (c)

(d) (e) (f)

Fig. 5: Qualitative analysis on the Historical Japanese Dataset by the instance-level segmentation model

5.2 Results on HJDataset

The results after training and evaluating our proposed instance-level segmentation model on the HJDataset are shown in Table 6. The mAP score has been computed in this case for seven different categories of objects (body, row, title, bio, name, position, and others) that have been detected and segmented as shown in the result table. Since the dataset contains hierarchical structured layouts as already shown in Figure 2 the task is quite challenging. A novel instance segmentation baseline has been provided for the evaluation of predicted mask instances by our proposed model. The overall Segmentation AP of 0.82 is quite good in such hierarchically variable layout elements. The results obtained for object detection task by our model have been compared with previous baselines provided by state-of-the-art models in Shen et. al. [45]. We achieve almost comparable state-of-the-art performance on detection with an overall AP of 0.822. Additionally, the AP scores obtained for individual object categories are quite comparable to the existing baselines on Faster-RCNN, RetinaNet and Mask-RCNN models.

Some interesting qualitative results for HJDataset has been shown in Figure 5. Figures 5(a), 5(b) and 5(f) clearly illustrate a lot of small objects which undergo dense overlapping with big objects. But the mask instances help these overlapped objects to be correctly

Table 6: Results for the HJDataset for the tasks of Document Object Detection and Document Instance Segmentation.

Category	Detection				Segmentation
	F-RCNN	M-RCNN	Retina	Ours	Ours
Body	0.990	0.991	0.990	0.992	0.996
Row	0.988	0.985	0.950	0.978	0.996
Title	0.876	0.895	0.696	0.891	0.913
Bio	0.945	0.868	0.895	0.937	0.944
Name	0.659	0.715	0.726	0.698	0.681
Position	0.841	0.842	0.859	0.862	0.862
Other	0.440	0.398	0.144	0.399	0.348
AP	0.819	0.813	0.752	0.822	0.820
AP@0.5	-	-	-	0.892	0.890
AP@0.75	-	-	-	0.876	0.878

predicted in the wild. The layout structure is extremely complex unlike the PubLayNet [61] dataset for scientific literature. The edges or boundaries of objects have a pixel-level classification that enhances improved predictions of our model. The 'name' category has the lowest AP in terms of both of both detection and segmentation performance, which is quite justified. It lies at an extremely low level in the hierarchy of layout objects in a page and has the smallest area. While the highest AP's for both detection and segmentation have been noted for 'Body' and 'Row' categories which belongs to the first and second level of hierarchy in the document structure. The model performance make sense for these large layout objects.

6 Conclusions

In this paper we have presented an instance segmentation model which, in contrast to classical document object detection pipelines, provides an evolution from coarse-grained detection with bounding boxes to a more fine-grained instance-level segmentation. Thus, it allows a deeper understanding of complex document layouts. To the best of our knowledge, this is the first work exploring instance segmentation techniques in the context of layout analysis and document understanding. This novel setting is especially important in document categories whose constituent blocks are not arranged in a tabular manner but with more complex layouts.

Furthermore, we have provided a comprehensive study showing the suitability of the proposed model on two business document datasets. In addition, our model demonstrates to be capable of dealing with several document typologies and scripts. In the tested settings, our proposed approach demonstrated a superior performance on both tasks, *i.e.* detection and segmentation.

There is a large future scope in the direction of instance-level segmentation for document layout understanding. This work provides a strong baseline for further improvement. More complex architecture designs have the potential to improve the performance of the proposed model baseline but it is not the key focus of this work. Also, there is further scope for improvement in the detection and segmentation of smaller objects.

Acknowledgements This work has been partially supported by the Spanish projects RTI2018-095645-B-C21, and FCT-19-15244, and the Catalan projects 2017-SGR-1783, the CERCA Program / Generalitat de Catalunya and PhD Scholarship from AGAUR (2021FIB-10010).

References

1. Agrawal M, Doermann D (2009) Voronoi++: A dynamic page segmentation approach based on voronoi and docstrum features. In: Proceedings of the International Conference on Document Analysis and Recognition, pp 1011–1015
2. Asi A, Cohen R, Kedem K, El-Sana J (2015) Simplifying the reading of historical manuscripts. In: Proceedings of the International Conference on Document Analysis and Recognition, IEEE, pp 826–830
3. Augusto Borges Oliveira D, Palhares Viana M (2017) Fast cnn-based document layout analysis. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp 1173–1180
4. Baechler M, Liwicki M, Ingold R (2013) Text line extraction using dmlp classifiers for historical manuscripts. In: 2013 12th International Conference on Document Analysis and Recognition, IEEE, pp 1029–1033
5. Binmakhashen GM, Mahmoud SA (2019) Document layout analysis: A comprehensive survey. ACM Computing Surveys 52(6):1–36
6. Cattoni R, Coianiz T, Messelodi S, Modena CM (1998) Geometric layout analysis techniques for document image understanding: a review. ITC-irst Technical Report 9703(09)
7. Chen J, Lopresti D (2011) Table detection in noisy off-line handwritten documents. In: Proceedings of the International Conference on Document Analysis and Recognition, pp 399–403
8. Dey S, Dutta A, Toledo JI, Ghosh SK, Lladós J, Pal U (2017) Signet: Convolutional siamese network for writer independent offline signature verification. arXiv preprint arXiv:170702131
9. Fang J, Gao L, Bai K, Qiu R, Tao X, Tang Z (2011) A table detection method for multipage pdf docu-

ments via visual seperators and tabular structures. In: Proceedings of the International Conference on Document Analysis and Recognition, pp 779–783
10. Gao L, Yi X, Jiang Z, Hao L, Tang Z (2017) Icdar2017 competition on page object detection. In: Proceedings of the International Conference on Document Analysis and Recognition, vol 1, pp 1417–1422
11. Gilani A, Qasim SR, Malik I, Shafait F (2017) Table detection using deep learning. In: Proceedings of the International Conference on Document Analysis and Recognition, vol 1, pp 771–776
12. Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1440–1448
13. Hao L, Gao L, Yi X, Tang Z (2016) A table detection method for pdf documents based on convolutional neural networks. In: Proceedings of the International Workshop on Document Analysis Systems, pp 287–292
14. He D, Cohen S, Price B, Kifer D, Giles CL (2017) Multi-scale multi-task fcn for semantic page segmentation and table detection. In: Proceedings of the International Conference on Document Analysis and Recognition, vol 1, pp 254–261
15. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778
16. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2961–2969
17. Huang Z, Huang L, Gong Y, Huang C, Wang X (2019) Mask scoring r-cnn. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 6409–6418
18. Isa D, Lee LH, Kallimani V, Rajkumar R (2008) Text document preprocessing with the bayes formula for classification using the support vector machine. IEEE Transactions on Knowledge and Data engineering 20(9):1264–1272
19. Jain A, Singh SK, Singh KP (2020) Handwritten signature verification using shallow convolutional neural network. Multimedia Tools and Applications pp 1–26
20. Journet N, Eglin V, Ramel JY, Mullot R (2005) Text/graphic labelling of ancient printed documents. In: Proceedings of the International Conference on Document Analysis and Recognition, pp 1010–1014
21. Kise K, Sato A, Iwata M (1998) Segmentation of page images using the area voronoi diagram. Computer Vision and Image Understanding 70(3):370–382
22. Koopman C, Wilhelm A (2020) The effect of preprocessing on short document clustering. Archives of Data Science, Series A 6(1):01
23. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp 1097–1105
24. Li J, Wei Y, Liang X, Dong J, Xu T, Feng J, Yan S (2016) Attentive contexts for object detection. IEEE Transactions on Multimedia 19(5):944–954
25. Li K, Wigington C, Tensmeyer C, Zhao H, Barmpalios N, Morariu VI, Manjunatha V, Sun T, Fu Y (2020) Cross-domain document object detection: Benchmark suite and method. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
26. Li XH, Yin F, Liu CL (2018) Page object detection from pdf document images by deep structured prediction and supervised clustering. In: Proceedings of the International Conference on Pattern Recognition, pp 3627–3632
27. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision, pp 740–755
28. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2117–2125
29. Lin TY, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2980–2988
30. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) Ssd: Single shot multibox detector. In: Proceedings of the European Conference on Computer Vision, pp 21–37
31. Marinai S, Gori M, Soda G (2005) Artificial neural networks for document analysis and recognition. IEEE Transactions on pattern analysis and machine intelligence 27(1):23–35
32. O'Gorman L (1992) The document spectrum for bottom-up page layout analysis. In: Advances in structural and syntactic pattern recognition, pp 270–279
33. Oliveira SA, Seguin B, Kaplan F (2018) dhsegment: A generic deep-learning approach for document segmentation. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR),

IEEE, pp 7–12
34. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al. (2019) Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, pp 8026–8037
35. Ramel JY, Leriche S, Demonet ML, Busson S (2007) User-driven page layout analysis of historical printed books. International Journal of Document Analysis and Recognition (IJDAR) 9(2-4):243–261
36. Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7263–7271
37. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp 91–99
38. Riba P, Dutta A, Goldmann L, Fornés A, Ramos O, Lladós J (2019) Table detection in invoice documents by graph neural networks. In: Proceedings of the International Conference on Document Analysis and Recognition
39. Roberts RJ (2001) Pubmed central: The genbank of the published literature
40. Saabni R, El-Sana J (2011) Language-independent text lines extraction using seam carving. In: Proceedings of the International Conference on Document Analysis and Recognition, IEEE, pp 563–568
41. Sahare P, Dhok SB (2019) Robust character segmentation and recognition schemes for multilingual indian document images. IETE Technical Review 36(2):209–222
42. Schreiber S, Agne S, Wolf I, Dengel A, Ahmed S (2017) Deepdesrt: Deep learning for detection and structure recognition of tables in document images. In: Proceedings of the International Conference on Document Analysis and Recognition, vol 1, pp 1162–1167
43. Sellen AJ, Harper RH (2003) The Myth of the Paperless Office. MIT Press, Cambridge, MA, USA
44. Shafait F, Smith R (2010) Table detection in heterogeneous documents. In: Proceedings of the International Workshop on Document Analysis Systems, pp 65–72
45. Shen Z, Zhang K, Dell M (2020) A large dataset of historical japanese documents with complex layouts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 548–549
46. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556
47. Soto C, Yoo S (2019) Visual detection with context for document layout analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing, pp 3455–3461
48. Staar PW, Dolfi M, Auer C, Bekas C (2018) Corpus conversion service: A machine learning platform to ingest documents at scale. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 774–782
49. Studer L, Alberti M, Pondenkandath V, Goktepe P, Kolonko T, Fischer A, Liwicki M, Ingold R (2019) A comprehensive study of imagenet pre-training for historical document image analysis. In: Proceedings of the International Conference on Document Analysis and Recognition, pp 720–725
50. Szegedy C, Ioffe S, Vanhoucke V, Alemi A (2016) Inception-v4, inception-resnet and the impact of residual connections on learning. arXiv preprint arXiv:160207261
51. Tran TA, Na IS, Kim SH (2015) Hybrid page segmentation using multilevel homogeneity structure. In: Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication, pp 1–6
52. Tupaj S, Shi Z, Chang CH, Alam H (1996) Extracting tabular information from text files. EECS Department, Tufts University, Medford, USA
53. Wei H, Baechler M, Slimane F, Ingold R (2013) Evaluation of svm, mlp and gmm classifiers for layout analysis of historical documents. In: 2013 12th International Conference on Document Analysis and Recognition, IEEE, pp 1220–1224
54. Wu Y, Kirillov A, Massa F, Lo WY, Girshick R (2019) Detectron2
55. Xiao Y, Yan H (2004) Location of title and author regions in document images based on the delaunay triangulation. Image and Vision Computing 22(4):319–329
56. Xie S, Girshick R, Dollár P, Tu Z, He K (2017) Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1492–1500
57. Xie Z, Huang Y, Jin L, Liu Y, Zhu Y, Gao L, Zhang X (2019) Weakly supervised precise segmentation for historical document images. Neurocomputing 350:271–281
58. Xu Y, Li M, Cui L, Huang S, Wei F, Zhou M (2020) Layoutlm: Pre-training of text and layout for document image understanding. In: Proceedings of the

26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 1192–1200
59. Yi X, Gao L, Liao Y, Zhang X, Liu R, Jiang Z (2017) Cnn based page object detection in document images. In: Proceedings of the International Conference on Document Analysis and Recognition, vol 1, pp 230–235
60. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Proceedings of the European Conference on Computer Vision, pp 818–833
61. Zhong X, Tang J, Yepes AJ (2019) Publaynet: largest dataset ever for document layout analysis. In: Proceedings of the International Conference on Document Analysis and Recognition, pp 1015–1022