Applied Intelligence
https://doi.org/10.1007/s10489-023-04782-3

HiM: hierarchical multimodal network for document layout analysis

Xu Canhui[1] · Li Yuteng[1] · Shi Cao[1] · Zhang Honghong[1] · Bi Hengyue[1] · Chen Yinong[2]

Accepted: 11 June 2023

Abstract
In document layout analysis, both computer vision based and natural language processing based methods are employed individually or integrated to enrich the feature information resources and to enforce object detection. To simultaneously leverage visual and textual modalities, this paper proposes a hierarchical multimodal (HiM) network to aggregate representative features from multi-source inputs with the introduction of complementary semantics and non-local context dependencies across grained scales. Different channel and spatial attention mechanisms are adapted to different modalities. The visual modality is based on conventional convolution network, while the textual modality focuses on embedding hierarchical textual vectors and positioning. The feature representations from multiple modalities are then integrated adaptively in feature pyramid network for subsequent region proposal processing. We have made database adaptation on PubLayNet, including inserting semi-structure elements and extending ground truth annotations by parsing PDF pages. On three popular benchmarks, including Article Regions, PubLayNet and DocBank, extensive experiments are carried out to verify the effectiveness and adaptability of HiM.

Keywords Multimodal · Object detection · Document layout analysis · Deep learning

1 Introduction

Document object detection is the task of automatically decomposing a document page image into it structural and logical units, such as figure, table, title, list, text, etc. It is provided critical foundation for a variety of document image analysis applications [1–3], such as document editing, document reflowable reading and information retrieval. Popular document formats, including page images (e.g., scanned and camera-captured documents) and digital born PDFs, do not explicitly encode document structure. To convert this document formats and extract the entire structure automatically has significant practical and academic values.

Although document understanding is a well-researched domain in past decades, document hierarchical structure recovery in various levels remains challenging. And multiple level such as character-level, line-level, fine-grained information are not fully exploited. Traditionally, multiple branches of resource information are aggregated intuitively, mainly including geometric, textual, typesetting and visual observation. However, the calculated features from various branches were still insufficient and rule-based heuristically [4]. Recently, there are various attempts in applying deep learning methods [5–11] to document images, including single stream computer vision (CV) methods, natural language processing (NLP) methods, multi-stream combining models and their varieties. CV-based models [12–18] identify regions with visual features, mostly deep convolution feature maps. The prominent capability of feature representation is one of the reasons why convolutional neural networks (CNN) [19] are popular in document detection. And NLP-based models [20–25] utilize sequential text labeling to group regions for layout analysis. These models employ textual information in form of word embedding.

✉ Shi Cao
caoshi@yeah.net

Xu Canhui
ccxu09@yeah.net

Li Yuteng
2420225023@qq.com

Zhang Honghong
may_i_zhh@163.com

Bi Hengyue
bihengyue@qq.com

Chen Yinong
yinong@asu.edu

[1] School of Information Science and Technology, Qingdao University of Science and Technology, 266000 Qingdao, ShanDong, China

[2] School of Computing and Augmented Intelligence, Arizona State University, Tempe, Arizona, USA

Published online: 23 July 2023

It is natural to integrate vision and textual information in multimodal, which conform to human reading process. Moreover, it is hypothesized that observations from different modalities are complementary for improving the performance. Therefore, approaches [23–29] combining both modalities are proposed for document object detection task. Specially, Yang et al. [26] first propose text embedding maps, which provide an innovative way to fully utilize textual information. VSR [28] is claimed to combine CV-based and NLP-based methods with the help of text embedding maps. Some multimodal architectures [26–28] integrate both textual information and visual information, others [23–25, 29] fuse both textual information and layout information. To explore the integration of these two modalities, numerous multi-models or multi-task document AI are based on Transformers. And the cross-modal alignment is critical for interaction of these two modalities. From our perspective, text modality can be fused into visual modality, and the interaction is expected to be learned by training.

Concerning hierarchical semantic structure recovery at different granularity, the intricate structure relation and context dependencies across grained scales remain unknown. To understand the context, various researches were carried out to tackle these unsolved problems. Conditional Random Fields based method managed to model the observation features and contexts [4]. Deep learning network architectures utilized attention subnetwork and graph convolution network modules [17, 30] for the purpose of context modeling. It could be observed that contextual clues in different granularity have their attributes. It should be discussed whether the added complementary contextual contribution is profitable for the current task. For multiple scales, non-local contextual information should be adjusted adaptively for our modalities.

To simultaneously leverage visual and textual modalities, in this paper, we propose a hierarchical multimodal (HiM) network to aggregate representative features from multi-source inputs with the introduction of complementary semantics and non-local context dependencies across grained scales. Particularly, textual modality focuses on embedding hierarchical textual vectors and positioning. The embedding operation is carried out in different granularities. Then, different channel and spatial attention mechanisms are adapted to different modalities. In addition, hierarchical evaluation dataset is introduced by inserting semi-structure elements and extending ground truth annotations by parsing PDF pages.

Our main contributions are summarized as follows:

- Hierarchical textual embeddings are obtained to describe textual information in multiple scales. Unlike numerous multi-models or multi-task based Transformers for document AI, textual modality is integrated into visual modality to learn cross-modal interaction.
- Attention mechanism is adapted to hierarchical level with the purpose of capturing non-local contextualization. Attention along channel and spatial dimensions is able to explore salient feature information.
- Standardized hierarchical evaluation dataset is possible to be obtained through our hierarchical dataset formation process. We evaluate HiM on three publicly available datasets. Extensive experimental results reveal that our proposed method can effectively boost document object detection performance.

2 Related works

The research of document object detection dates to the early 1990s. The earlier rule-based works [31–36] consist of bottom-up approaches and top-down approaches. Bottom-up approaches [31, 34, 36] first recognize individual characters, and then cluster them into tokens, lines, and zones to form the structure of the entire document. The top-down approaches [32, 33] usually recursively segment a document page into columns, blocks, text lines, and words. However, these methods are of limited value in correctly segmenting document with complex layout structure, such as irregular figures, tables and lists in a document [37]. Early conventional machine learning approaches [35, 38] had been widely utilized for document detection tasks. Support Vector Machines, Multi-Layer Perceptrons and Gaussian Mixture Models classifiers [38] have been evaluated in document object detection research. However, these methods are restricted to low dimensional handcrafted features, which are hard to be designed and prone to errors.

Early deep learning approaches mainly exploited single stream visual modality in document-related tasks [12–16, 39]. Various backbone architectures such as YOLO [12], Faster R-CNN [13], Mask R-CNN [14] and Cascade R-CNN [15] have been proposed to utilize deep convolution feature maps to detect specific document objects. These task oriented page object detectors have made progress in terms of table [40], figure [41], text block and line [42]. Agarwal et al. [43] employ an end-to-end trainable deep network CDeC-Net to detect tables, based on Cascade R-CNN [15] with a dual backbone. The method can obtain a high detection result at higher intersection over union (IoU) threshold. Paliwal et al. [44] introduce TableNet for both table detection and structure recognition. The object regions are split by the interdependence table detection and table structure recognition. And another single stream textual modality focuses on NLP-based methods [20–23] by utilizing sequential text labeling. It is suggested that complementary information from different modalities could be integrated wholistically instaed of restricting to single stream information.

Multimodal approaches [26–28, 45–48] attempt to combine multiple stream modality information, including textual, visual, style and layout information, which could be roughly categorized into two groups: shallow modality fusion and deep modality fusion methods [29]. The former one individually processes textual information extracted by pretrained NLP-based models and visual information learned from image-centric frameworks, then integrates both modality features. Yang et al. [26] utilize visual information and textual information by introducing text embedding maps to make use of textual representation. Barman et al. [27] directly integrate pixel-level visual features and text embedding maps for historical newspaper layout analysis tasks. Zhang et al. [28] propose a two-stream network by extracting visual and semantic features. Then the features are adaptively aggregated to learn model realtions. However, shallow-based approaches might have a limited performance to be adapted into document scenarios in type diversity.

The latter one, named deep fusion-based methods [24, 25, 29, 45–49], try to robustly learn multimodal information by applying large pre-training strategies on a great number of unlabeled datasets. Compared to shallow fusion-based methods, deep fusion-based methods not only can be beneficial from modality information extraction, but also learns to bridge cross-modal interaction in different domains. LayoutLM [24] is one of the very first attempts to design a single framework for both textual and layout feature learning from text-centric document images. Semantically meaningful text embeddings and visually rich visual embeddings are aligned into the BERT [21] model. After model pre-training, LayoutLM could be adopted as a basis backbone for various downstream tasks of document AI, which brings richer semantic and visual representations. LayoutLMv2 [29] is proposed to jointly model textual, visual and layout information, and takes advantages to bridge cross-modal interaction among three modalities for effective text recognition. A novel pre-training technique is designed for LayoutLMv2 training, including text-image alignment and matching. The applied Transformer-based model takes text, layout and visual embeddings as input to establish cross-modal correlation and interaction. LayoutLMv3 [49] is a simple yet efficient unified multimodal framework for both text-centric and image-centric Document AI applications. LayoutLMv3 [49] is the first attempt to extract image embeddings without CNNs, which greatly reduces time and space complexity. To overcome the cross-modal problem, LayoutLMv3 [49] is pre-trained with Masked Language Modeling, Masked Image Modeling and Word-Patch Alignment objectives in a self-supervised way, being capable of better representation learning between different modalities.

For document images, there exists a great number of intrinsic and inherent structured relation and contextual information across different grained scales. Text fileds of interest present the human reading order. For example, figure captions as the depictions of pictures always appear in the below of figures, which is convenient for human understanding. Text fields usually carry different granularities and valuable clues. An annotated region could be segmented into character-level, word-level and sentence or instance-level. Learning representation from different levels might be an optimal way for high-level document understanding. Inspired from the above-mentioned methods, to simultaneously leverage visual and textual modalities, we propose the HiM network to aggregate representative features from multi-source inputs with the introduction of complementary semantics and non-local context dependencies across grained scales.

3 Hierarchical multimodal network

The architecture of our proposed method HiM network is illustrated in Fig. 1, mainly including three parts: visual convolution module, hierarchical textual embedding with attention module, feature pyramid fusion module. For visual modality, regional deep convolution features are extracted due to prominent representation capabilities. For textual modality, complementary semantics and non-local context dependencies are introduced across multiple grained scales. Hierarchical textual vectors and positioning are embedded with different channel and spatial attention mechanisms. To unify visual and textual modalities, bottom-up feature pyramid introduces lower feature maps with larger spatial information to topmost layer with rich semantic details, while well-known top-down feature pyramid network (FPN) [50] is upsampled to propagate deeper high-level semantic information to shallower layers so as to enhance performance semantically. The fused features are fed into region proposal network (RPN) to extract a set of rectangular object proposals and improve the speed of predicting bounding box results.

3.1 Hierarchical text embedding maps

To jointly utilize multimodal input resource, our proposed method HiM attempts to extract the visual and textual information. Particularly, textual modality could be fused into visual modality. With the parsed semi-structure positioning information or OCRed outputting coordinates, the bounding box information can be used to convert the textual information in two-dimensional space visually [26–28].

From PDF parsed XML files or OCRed (Optical Character Recognition) outputs, bounding boxes at multiple scales are denoted as $b = (x_0, y_0, x_1, y_1)$, including character level box set b^c, line fragment level box set b^f and block level box set b^b. Thus, given multiple levels $l = \{0, 1, \cdots, n\}$, $S^l = \{S^0, S^1, \cdots, S^n\}$ is denoted as hierarchical comple-

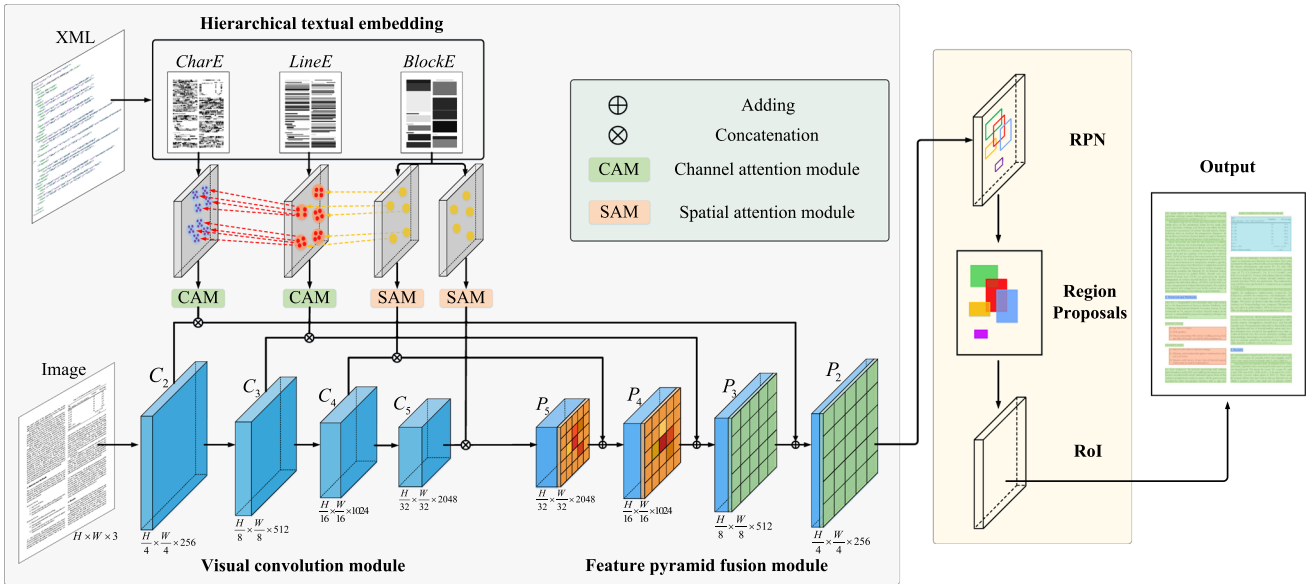

Fig. 1 The architecture of the proposed HiM network mainly contains three parts: visual convolution module, hierarchical textual embedding with attention module, feature pyramid fusion module

mentary structural information. S_k and b_k are related to k-th bounding box at different levels. In a document image $I \in \mathbb{R}^{H \times W}$, each pixel $I_{i,j}$ belonging to b at different scale, shares the same embedding vector at its corresponding level. Specifically, hierarchical textual embedding maps are formulated as follows:

$$CharE_{i,j} = \begin{cases} E^c(S_k^0) & if(i,j) \in b_k^c \\ 0 & otherwise \end{cases} \tag{1}$$

$$LineE_{i,j} = \begin{cases} E^f(S_k^1) & if(i,j) \in b_k^f \\ 0 & otherwise \end{cases} \tag{2}$$

$$BlockE_{i,j} = \begin{cases} E^b(S_k^2) & if(i,j) \in b_k^b \\ 0 & otherwise \end{cases} \tag{3}$$

where E is a mapping function converting textual information to vectors, and zero vector represents pixel containing no text. As can be seen in Fig. 1, hierarchical textual embeddings, including $CharE$, $LineE$, and $BlockE$, are thus obtained. Subsequently, hierarchical textual embedding maps can be processed directly by conventional neural convolution to construct page objects features observation $T^{S^l} = \{T_i^{S^l}\}$. As for the channel dimension of embedded feature representation, the experimental section has presented detailed discussion.

3.2 Channel and spatial attention mechanism

Note that our proposed method HiM further processed the extracted hierarchical embedding feature maps, instead of simply concatenating the same channel dimension of textual embedding maps. It is observed that lower-level features depend less on context than deeper layers with rich semantic information. Fine-grained segments generally contain excessive contexts which are less likely to contributes to large object detection. On the contrary, coarse-grained segments could put more emphasis on profitable context for deeper layer semantic information. Thus, we apply different attention mechanisms adapted to different modalities: channel attention module (CAM) and spatial attention module (SAM). For the lower character and line level textual feature representation, Squeeze-and-Excitation (SE) [51] is applied in CAM to emphasize channel attention, while on the other hand, the observation inspires us to apply deformable convolution [52] in SAM to embeddings in block level.

The low-level feature maps focus on the prediction of small objects which depend less on contextual information. The prediction of common objects like tables, figures and lists needs more contextual information in high-level feature maps. These objects include more contents such as table caption, figure caption and the symbol of lists. Normal convolution is unable to explore more contextual information due to its fixed receptive field. The deformable convolution can provide more flexible receptive fields because the sampling locations of its convolution kernel is irregular. The learning offsets can augment the spatial sampling locations to acquire more information related to prediction objects.

As shown in Fig. 1, SAM is applied in higher stages to extract spatial feature representations. With the extracted embedding feature representations T^{S^l}, deformable convolution in SAM is defined as:

$$(\mathrm{K} * \mathrm{T})(x) = \sum_{p_i \in P} \mathrm{K}(p_i)\mathrm{T}(x + p_i + \Delta d) \tag{4}$$

where K denotes a 3×3 kernal with dilation 1, P defines the receptive field size, and p_i enumerates the locations in P. T represents input feature map, x defines a location on the output feature map, and Δd is the offset which is typically fractional.

The feature representations with attention from visual and textual modalities are then integrated adaptively in FPN for subsequent region proposal processing. The increased lateral attention modules are expected to bring up additional textual structure information precision. The output FPN aggregated the lateral attentional connections, which is formulated as:

$$\begin{cases} P_5 = Concat(Conv_{C_5}(C_5), D(T^{S^2})) \\ P_4 = F(Conv_{P_5}(P_5)) + Concat(Conv_{C_4}(C_4), D(T^{S^2})) \\ P_m = F(Conv_{P_{m+1}}(P_{m+1})) + Concat(Conv_{C_m}(C_m), \\ \qquad SE(T^{S^{m-2}})) \end{cases} \tag{5}$$

where $Concat$ denotes the concatenation, $Conv_C$ and $Conv_P$ are different convolution operations. T^S represents the hierarchical text embedding map. F is the upsampling operation. D is deformable convolution operation. SE is the Squeeze-and-Excitation module [51] and $m \in \{2, 3\}$.

3.3 Document hierarchical structure formation

To our knowledge, standardized hierarchical evaluation dataset is rarely available. Little effort has been made on large public hierarchical document datasets so as to achieve document structure recovery holistically. With hierarchical document structure dataset, the semantic structure within page and across multiple pages could be learned through deep learning network, with the aim of facilitating the usability of recovery multi-scale structure. Multiple level document structure requires segmentation in different granularities, ranging from coarse to fine grained level. Figure 2 illustrates a visual description of document hierarchy in a sample page. Block segments at high level are rough paragraphs with clusters of line segments at low level. In this way, higher level segments aggregate lower level homogeneous primitives. Partial compoments and wholistical objects could bring complementary information for multi-scale segmentation.

In order to explore hierarchical relation within a document, segmentation results at multiple levels are considered, ranging from basic page primitives at low level to regional grouping blocks at top level. With effective parser such as PyMuPDF, page content stream can be extracted, including text, images and graphics at different levels. The general document structure consists of "chars", "words", "spans", "blocks", where blocks are roughly paragraphs with clusters of lines, a line consists of spans, and a span includes characters with identical formatting properties such as fonts. PDF parser is capable of providing ground truth in different granularities. Given multiple levels $l = \{0, 1, \cdots, n\}$, hierarchical segments $S^l = \{S^0, S^1, \cdots, S^n\}$, and page objects features observation $X^{S^l} = \{X_i^{S^l}\}$, with corresponding semantic labels $Y^{S^l} = \{Y_i^{S^l}\}$. Higher level segments are composed by lower-level divisions. For each level S^l, observation $X_i^{S^l}$ is composed of lower level $X_i^{S^{l-1}}$.

As shown in Fig. 3, the XML schema contains three level segmentation elements, represented by different background colors. Each segmentation element in S^l can be symbolized as a circle. Segments at higher levels aggregate clusters of lower-level components. For $l = 2$, S^{l-2} level segment includes meta information of primitive contents like characters derived from PDF files. S^{l-1} represents line level or fragments aggregating homogeneous primitives with proximity and formatting properties. S^l denotes block level with an aggregation of line segments. Precise identifiers within ground truth have advantage of alleviating the problem of overlapping regions in documents with complex layouts.

4 Experimental results and discussion

To comprehensively evaluate the performance of our network HiM, a series of experiments are conducted on three public document datasets, including Article Regions, PubLayNet and DocBank.

4.1 Datasets

Article Regions contains 822 document images generated from 100 scientific articles sampled by PMC's open access collection, including 9 categories (Title, Authors, Abstract, Body, Figure, Figure Caption, Table, Table Caption, and References). The default format is PASCAL VOC, and we convert it into COCO format. The dataset is split into 600 and 222 images for training and test. And the mean average precision (mAP) [53] is exploited as the evaluation metric by default.

PubLayNet is a large dataset of document images, of which the layout is annotated with both bounding boxes and polygonal segmentations. The dataset is utilized in ICDAR 2021 competition Scientific Literature Parsing (SLP) task A [54].

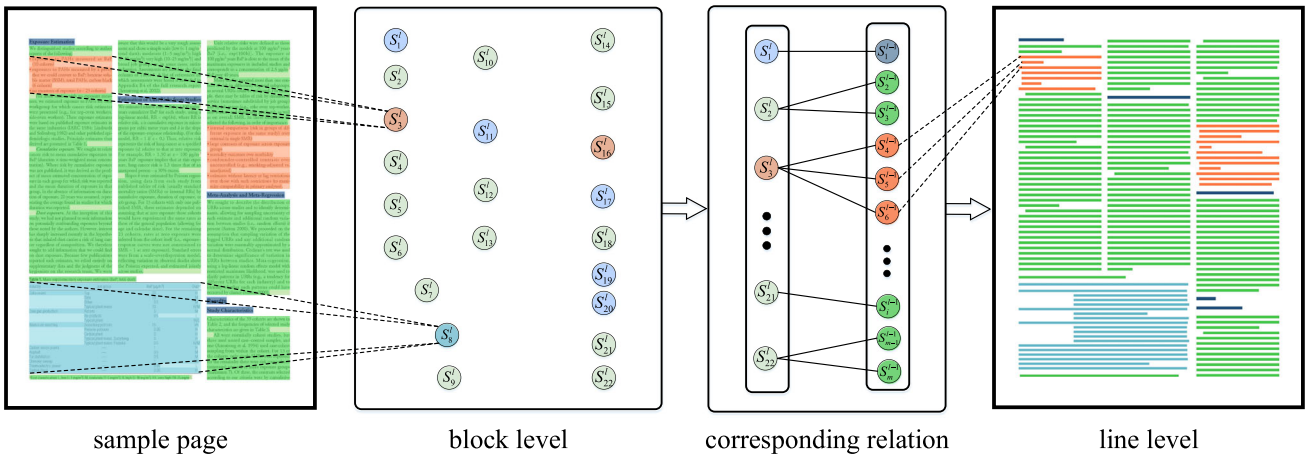

Fig. 2 Visual description of document hierarchy of a sample page. Block segments at higher level aggregate clusters of line segments at lower level

It contains more than 360,000 document images, with 5 categories: Text, Title, List, Table, and Figure. Specifically, the official splits contain 335,703 training images, 11,245 validation images, and 11,405 test images. It should be noted that PubLayNet dataset provides only block level ground truth, while our adapted dataset provides parsed semi-structure annotation additionally, described by XML format. We utilize overall mAP at IoU [0.5:0.95] as the evaluation metric.

DocBank is a large document-level benchmark proposed by Microsoft. It consists of 500,000 document pages with token-level annotations for layout analysis, including 12 categories (Abstract, Author, Caption, Equation, Figure, Footer, List, Paragraph, Reference, Section, Table, and Title). The dataset contains 400,000 training images, 50,000 validation images, and 50,000 test images. F1 score is used as official evaluation metric.

4.2 Implementation details

Our network HiM is implemented on PyTorch framework. It is trained on one TITAN Xp GPU by the stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and weight decay of 0.0001. The batch size is set to 2, and each image has 512 sampled RoIs with a ratio of 1:3 of positive-negative. ResNeXt-101 is utilized as backbone network. On PubLayNet and DocBank datasets, it is trained 6 epochs at a starting learning rate of 0.0009, with decay of 0.1 every 2

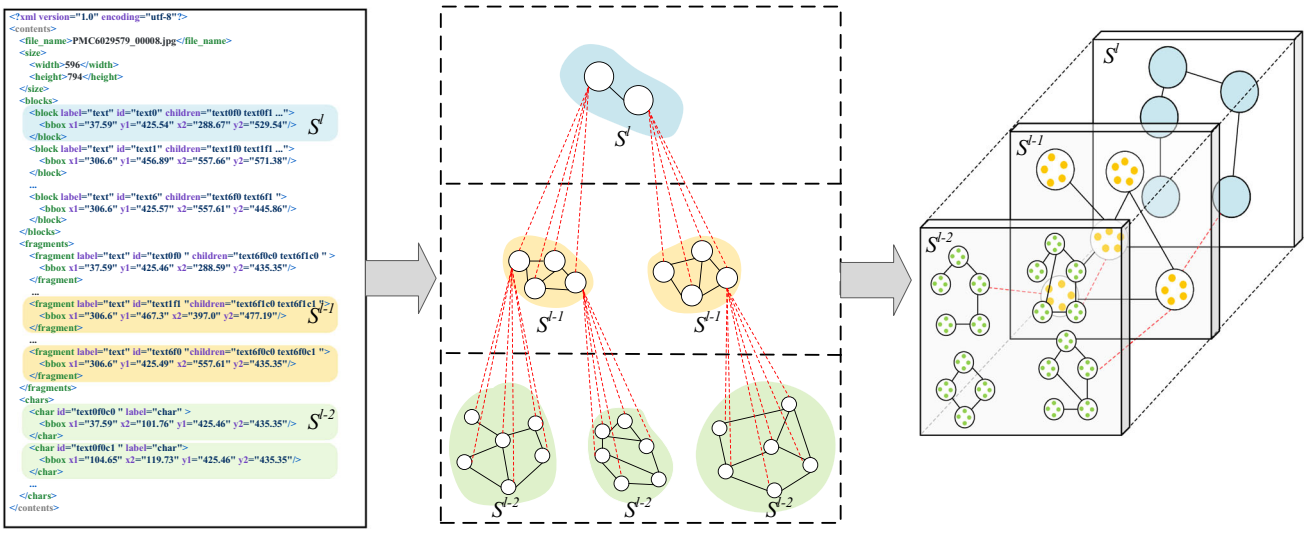

Fig. 3 The XML schema is given for describing hierarchical segments, blocks and logical labels. For $l = 2$, S^{l-2} level segment includes meta information of primitive contents like characters derived from PDF files. S^{l-1} represents line level or fragments aggregating homogeneous primitives with proximity and formatting properties. S^l denotes block level with an aggregation of line segments. Segments at higher levels are made of their lower-level components by using the identifiers

epochs. The training on Article Regions takes 30 epochs at a starting learning rate of 0.0025, with decay of 0.1 every 10 epochs. All of the experiments are conducted on a workstation with Ubuntu 16.04 and an Intel(R) Xeon(R) E5-2650 v4 2.20GHz CPU 251GB RAM.

4.3 Experimental results

PubLayNet Table 1 shows the performance comparison of HiM and seven methods on PubLayNet validation dataset. Our model HiM achieves 95.09% mAP, which surpasses all the comparable methods except VSR, with a slight decrease 0.6% mAP. But HiM is superior to VSR on the List, Table and Figure categories. It is clear that Faster R-CNN, Mask R-CNN, ATSS [55] and CDeCNet only focus on extracting visual modality information. HiM has its superiority over single stream unimodal methods, which shows similar trend of performance improvement for other multimodal methods, such as VSR, DiT, and LayoutLMv3. Among the multimodals, VSR, as the winner of ICDAR 2021 competition SLP task A, integrates visual feature, textual feature and component relation. It outperforms all the methods with overall 95.66% mAP. But it is inferior to HiM on the List with a decrease of 1.03%, Table with 0.51% and Figure with 1.12%. Our model HiM utilizes hierarchical information of multiple modalities in visual and textual streams. DiT [56] and LayoutLMv3 [49] unify visual and textual modalities by Transformer. We compare multimodal Transformer with our HiM. HiM achieves a high gain in almost all of categories, including Text, List, Table, and Figure. Since the small bullets or numbers of List are typically much smaller, fine grained textual embedding could capture the details. LayoutLMv3 points out that it is unnecessary to extract image features by CNN, while our model HiM advocates that text modality can be fused into visual modality.

Figure 4 shows some visual comparison samples of page object detection on PubLayNet. It is observed that an obvious overlapping green bounding box across three columns in Fig. 4(b-3), is detected by Faster R-CNN. In Fig. 4(b-4), it can be spotted that imprecise boxes for Table and Figure categories are overlaid around boundaries. As shown in Fig. 4(b-2), it calibrates the inaccurate prediction box and identifies the right logical roles of the region. In our model HiM, the overlapping issue is alleviated. In this case, HiM obtains higher precision of box regression with the contribution of extra textual modality information. Interestingly, Fig. 4(c-3) and (c-4) misclassify the orange List to green Text. It is admitted that text-related classes(Text, Title, and List) are difficult to distinguish. Our HiM detects the long List correctly since it utilizes hierarchical semi-structure textual embeddings, especially the low level List symbols information. In Fig. 4 (a-3), (a-4) (d-3) and (d-4), pink Title is misclassified as green Text, but HiM is able to detect correctly.

Article Regions In Table 2, our model is compared with four different methods [13, 14, 28, 57] on Article Regions dataset. From experimental results, our model HiM surpasses all methods, showing the highest overall mAP with 95.92%. HiM outperforms on Table caption, with 5.7% increase on AP and on Figure caption with 3.1% when compared with VSR. Table caption and Figure caption regions are relatively small object, which could be neglected via CV-based methods. Complementary textual embeddings can provide primitive information for small objects. Our model HiM is capable of recognizing these regions precisely. Experimental results demonstrate that visual feature and textual feature could leverage complementary information and help improve detection accuracy.

DocBank We evaluate the document layout task on DocBank [58] dataset since it has 12 categories and a variety of complex layouts. In Table 3, we compare HiM with BERT, RoBERTa, LayoutLM [24], Faster R-CNN with ResNeXt-101, ensemble models (ResNeXt-101+LayoutLM), and VSR. It is observed that HiM obtains the highest F1 scores on most categories, including Author, Caption, Equation, Footer, Reference, Section, Table and Title. Particularly, HiM outperforms VSR on Table with 3.14%. The results indicate that HiM is capable of achieving better performance on DocBank with fine grained categories and complex layouts.

Table 1 Performance comparison on PubLayNet validation dataset

Method	Text	Title	List	Table	Figure	mAP
Faster R-CNN [13] (2016)	91.73	82.66	88.27	95.46	93.73	90.37
Mask R-CNN [14] (2018)	91.71	84.21	88.77	96.23	95.11	91.13
ATSS [55] (2020)	93.68	83.69	93.83	96.56	95.17	92.59
CDeCNet [43] (2021)	91.30	84.20	89.70	96.70	90.50	90.48
VSR [28] (2021)	**96.70**	**93.10**	94.70	97.40	96.40	**95.66**
DiT$_{base}$ [56] (2022)	94.40	88.90	94.80	97.60	96.90	94.50
LayoutLMv3$_{base}$ [49] (2022)	94.50	90.60	95.50	97.90	97.00	95.07
HiM	94.51	89.77	**95.73**	**97.91**	**97.52**	95.09

Fig. 4 Qualitative results of the proposed HiM and the compared methods on PubLayNet

Table 2 Performance comparison on Article Regions test dataset

Method	Title	Author	Abstract	Body	Figure	Figure Caption	Table	Table	References	mAP
Faster R-CNN w/context [13]	–	10.34	–	93.58	–	–	–	30.80	–	70.30
Faster R-CNN [13]	92.71	79.90	93.79	96.83	86.87	92.66	81.66	68.83	98.34	87.95
Mask R-CNN [14]	83.85	78.05	94.20	97.42	90.43	93.78	83.39	71.64	99.37	88.01
CDDOD [57]	88.61	63.24	87.61	93.35	82.92	87.94	82.36	69.64	94.30	83.33
VSR [28]	**100.00**	**94.00**	95.00	99.10	95.30	94.50	**96.10**	84.60	92.30	94.50
HiM	97.90	90.80	**96.50**	**99.70**	**95.40**	**97.60**	95.10	**90.30**	**100.00**	**95.92**

4.4 Ablation studies

In this section, we systematically evaluate the effectiveness of each module in HiM, the impact of the modules fusion strategies and integrated textual embedding layer configurations. And a set of experiments are conducted on PubLayNet dataset. Especially, each model is trained for 90k iterations with a learning rate of 0.0025. The baseline model is implemented by Faster R-CNN with adapted text bounding box regression branch.

Effectiveness of Hierarchical Textual Embedding To verify the promotion of the hierarchical textual embeddings, we

Table 3 Performance comparison on DocBank test dataset in F1 Score

Method	Abstract	Author	Caption	Equation	Figure	Footer	List
BERT_{base}	92.94	84.84	86.29	81.52	100.00	78.05	71.33
RoBERTa_{base}	92.88	86.18	89.44	82.48	100.00	80.14	73.53
LayoutLM_{base}	98.16	85.95	95.97	89.47	100.00	89.57	89.48
BERT_{large}	92.86	85.77	86.50	81.77	100.00	78.14	69.60
RoBERTa_{large}	94.79	87.24	90.81	83.70	100.00	83.92	74.51
LayoutLM_{large}	97.84	87.83	95.56	89.74	**100.00**	91.46	90.04
Faster R-CNN+X101	97.17	82.27	94.35	89.38	88.12	90.29	90.51
$\text{X101+LayoutLM}_{base}$	98.15	89.07	96.69	94.30	99.90	92.92	93.00
$\text{X101+LayoutLM}_{large}$	98.02	89.64	96.66	94.40	99.94	93.52	92.93
VSR	**98.29**	91.19	96.32	95.84	99.96	95.11	**94.66**
HiM	96.37	**92.08**	**97.18**	**96.52**	98.44	**96.39**	92.72

Method	Paragraph	Reference	Section	Table	Title	Macro Average
BERT_{base}	96.19	93.10	90.81	82.96	94.42	87.70
RoBERTa_{base}	96.46	93.41	93.37	83.89	95.11	88.91
LayoutLM_{base}	97.88	93.38	95.98	86.33	95.79	93.16
BERT_{large}	96.19	92.84	90.65	83.20	94.30	87.65
RoBERTa_{large}	96.65	93.34	94.07	84.94	94.61	89.88
LayoutLM_{large}	97.90	93.32	95.96	86.79	95.52	93.50
Faster R-CNN+X101	96.82	87.98	94.12	83.53	91.58	90.51
$\text{X101+LayoutLM}_{base}$	98.43	94.37	96.64	88.18	95.75	94.78
$\text{X101+LayoutLM}_{large}$	98.44	94.30	96.70	88.75	95.31	94.88
VSR	**98.66**	95.05	97.11	89.24	95.63	95.59
HiM	96.59	**95.56**	**97.61**	**92.38**	**97.15**	**95.75**

Table 4 Ablation study on PubLayNet benchmark with different modules. The compared modules are denoted as follows: Hierarchical Textual Embeddings, CAM, and SAM

	Hi	CAM	SAM	Text	Title	List	Table	Figure	mAP	time(ms)
Baseline	–	–	–	91.83	82.53	84.40	95.68	91.58	89.20	73.00
+Hi	√	–	–	91.76	83.70	**90.09**	96.73	94.06	91.27	90.70
+CAM	–	√	–	91.73	83.11	88.02	96.33	94.00	90.64	146.60
+SAM	–	–	√	91.54	83.29	88.09	96.59	94.37	90.78	160.00
HiM	√	√	√	**91.84**	**84.12**	89.82	**97.64**	**95.07**	**91.70**	187.00

conduct experiments to show how the baseline could be improved if aggregation of line fragments and char fragments embeddings is considered. Table 4 shows experimental results on PubLayNet [59] dataset. Obviously, when the auxiliary branch hierarchical textual features are equipped with baseline, the performance is boosted from 89.20% to 91.27% mAP at a small amount of additional computational cost. The result is 5.69% AP higher on List and 2.48% AP higher on Figure than baseline. By merging both information from visual and textual modalities, the feature representations are further enhanced. It indicates that detection performance can benefit from auxiliary input source to enrich the feature representation. And it might attribute to the complementary features representing visual information and fine-grained layout details.

Effectiveness of CAM and SAM Objects like title, is identified with the embedding at character level. Since it depends less on contextual information, a SE attention module [51] is applied thereby. When equipped with CAM alone, the result increases to 90.64% mAP and outperforms more in Title, List and Figure categories. CAM is able to effectively emphasize critical features, such as small bullets or numbers of list and borders of Table or Figure in channel dimension. And this further demonstrates the importance of textual information for helping improve object detection accuracy. Then, the baseline is equipped with deformable convolution module in SAM. From the results, it performs higher mAP than baseline model, which boosts from 89.20% to 90.78%. Deformable convolution is capable of guiding the model to locate the informative regions by learning spatial offset. Therefore, more spatial contextual information can be captured. The features at higher level are profitable for contextual objects like table and table caption or figure and figure caption.

Effectiveness of integrated textual embedding layers To leverage the feature representation from both text modality and image modality, the ratio of embedded hierarchical layers T to visual feature layers V, denoted as σ, is used to depict different layer combinations. Table 5 shows experimental results on PubLayNet dataset. There listed six types of ratio configuration for integration of these two modalities. Ratio σ with 1:15 achieves highest improvement by around 2.07% at a small amount of additional computational cost. When integrating limited textual embedding layers, it produces insignificant impact for feature representation. On the contrary, adding equal number of textual embedding layers over visual layers might result in feature confusion effects.

Above all, HiM integrates these three modules, including Hierarchical Textual Embedding, CAM and SAM with reasonable layer ratio, which enables the result achieving superior performance with 91.70%. It could be noted that the complementary feature representations with multiple level fine-grained textual attributes play a significant role for detection performance. Furthermore, CAM and SAM are able to enhance smaller regions feature, which attributes to more contextual information and selection of critical features.

5 Conclusion

In this paper, we introduced a simple yet effective hierarchical multimodal network HiM with channel and spatial attention for document layout analysis, which simultaneously exploited both visual and textual information. To aggregate representative features from multi-source inputs and introduce complementary semantics, hierarchical textual vectors and positioning was embedded in different granularities. Dif-

Table 5 Ablation study on PubLayNet benchmark with different integrated textual embedding layer ratio configuration

T-V-ratio	Text	Title	List	Table	Figure	mAP	time(ms)
0:1	91.83	82.53	84.40	95.68	91.58	89.20	73.00
1:1	91.58	83.34	89.94	96.30	94.25	91.08	89.70
1:3	**91.90**	83.41	89.85	96.59	**94.26**	91.20	89.00
1:7	91.73	83.56	**90.40**	96.61	93.87	91.23	90.60
1:15	91.76	**83.70**	90.09	**96.73**	94.06	**91.27**	90.70
1:31	91.58	83.68	90.10	96.59	94.16	91.22	91.30

ferent channel and spatial attention mechanism are adapted to textual modality. The feature representations from multiple modalities were then integrated adaptively in FPN for subsequent region proposal processing. On three public document datasets, including Article Regions, PubLayNet and DocBank the experimental results demonstrate the superiority and generalization ability of HiM.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant No. 61806107 and 61702135, Shandong Key Laboratory of Wisdom Mine Information Technology, and the Opening Project of State Key Laboratory of Digital Publishing Technology.

Data Availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflicts of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Liao M, Zou Z, Wan Z, Yao C, Bai X (2023) Real-time scene text detection with differentiable binarization and adaptive scale fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(1):919–931. https://doi.org/10.1109/TPAMI.2022.3155612
2. Xu C-h, Shi C, Chen Y-n (2021) End-to-end dilated convolution network for document image semantic segmentation. Journal of Central South University 28(6):1765–1774
3. Xu C, Shi C, Bi H, Liu C, Yuan Y, Guo H, Chen Y (2021) A page object detection method based on mask r-cnn. IEEE Access 9:143448–143457. https://doi.org/10.1109/ACCESS.2021.3121152
4. Tao X, Tang Z, Xu C (2014) Contextual modeling for logical labeling of pdf documents. Computers & Electrical Engineering 40(4):1363–1375
5. Liao M, Lyu P, He M, Yao C, Wu W, Bai X (2021) Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence 43(2):532–548. https://doi.org/10.1109/TPAMI.2019.2937086
6. Xu Y, Fu M, Wang Q, Wang Y, Chen K, Xia G-S, Bai X (2020) Gliding vertex on the horizontal bounding box for multi-oriented object detection. IEEE transactions on pattern analysis and machine intelligence 43(4):1452–1459
7. Shi B, Yang M, Wang X, Lyu P, Yao C, Bai X (2019) Aster: An attentional scene text recognizer with flexible rectification. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(9):2035–2048. https://doi.org/10.1109/TPAMI.2018.2848939
8. Cai Y, Liu Y, Shen C, Jin L, Li Y, Ergu D (2022) Arbitrarily shaped scene text detection with dynamic convolution. Pattern Recognition 127:108608
9. Liu Y, Shen C, Jin L, He T, Chen P, Liu C, Chen H (2022) Abcnet v2: Adaptive bezier-curve network for real-time end-to-end text spotting. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(11):8048–8064. https://doi.org/10.1109/TPAMI.2021.3107437
10. Liu C, Liu Y, Jin L, Zhang S, Luo C, Wang Y (2020) Erasenet: End-to-end text removal in the wild. IEEE Transactions on Image Processing 29:8760–8775
11. Zhang, H, Yao, Q, Kwok, J.T, Bai, X (2022) Searching a high performance feature extractor for text recognition network. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp 1–15. https://doi.org/10.1109/TPAMI.2022.3205748
12. Redmon, J, Divvala, S, Girshick, R, Farhadi, A (2016) You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788
13. Ren S, He K, Girshick R, Sun J (2017) Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(6):1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031
14. 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
15. Cai Z, Vasconcelos N (2019) Cascade r-cnn: high quality object detection and instance segmentation. IEEE transactions on pattern analysis and machine intelligence 43(5):1483–1498
16. 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
17. Bi, H, Xu, C, Shi, C, Liu, G, Li, Y, Zhang, H, Qu, J (2022) Srrv: A novel document object detector based on spatial-related relation and vision. IEEE Transactions on Multimedia, pp 1–1. https://doi.org/10.1109/TMM.2022.3165717
18. Shi C, Xu C, Bi H, Cheng Y, Li Y, Zhang H (2022) Lateral feature enhancement network for page object detection. IEEE Transactions on Instrumentation and Measurement 71:1–10. https://doi.org/10.1109/TIM.2022.3201546
19. LeCun, Y, Bengio, Y, et al. (1995) Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks **3361**(10):1995
20. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A et al (2020) Language models are few-shot learners. Advances in neural information processing systems 33:1877–1901
21. Devlin, J, Chang, M-W, Lee, K, Toutanova, K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, vol 1 (Long and Short Papers), pp 4171–4186. Association for Computational Linguistics,Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-1423. https://aclanthology.org/N19-1423
22. Liu, Y, Ott, M, Goyal, N, Du, J, Joshi, M, Chen, D, Levy, O, Lewis, M, Zettlemoyer, L, Stoyanov, V (2019) Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692
23. Garncarek, Ł, Powalski, R, Stanisławek, T, Topolski, B, Halama, P, Turski, M, Graliński, F (2021) Lambert: Layout-aware language modeling for information extraction. In: International conference on document analysis and recognition, pp 532–547. Springer
24. 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
25. Xu, Y, Lv, T, Cui, L, Wang, G, Lu, Y, Florencio, D, Zhang, C, Wei, F (2021) Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. arXiv:2104.08836
26. Yang, X, Yumer, E, Asente, P, Kraley, M, Kifer, D, Lee Giles, C (2017) Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In: Pro-

ceedings of the IEEE conference on computer vision and pattern recognition, pp 5315–5324
27. Kaplan, F, Oliveira, SA, Clematide, S, Ehrmann, M, Barman, R (2021) Combining visual and textual features for semantic segmentation of historical newspapers. Journal of Data Mining & Digital Humanities
28. Zhang, P, Li, C, Qiao, L, Cheng, Z, Pu, S, Niu, Y, Wu, F (2021) Vsr: A unified framework for document layout analysis combining vision, semantics and relations. In: International conference on document analysis and recognition, pp 115–130. Springer
29. Xu, Y, Xu, Y, Lv, T, Cui, L, Wei, F, Wang, G, Lu, Y, Florencio, D, Zhang, C, Che, W, Zhang, M, Zhou, L (2021) LayoutLMv2: Multimodal pre-training for visually-rich document understanding. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (vol 1: Long Papers), pp 2579–2591. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2021.acl-long.201. https://aclanthology.org/2021.acl-long.201
30. Shi C, Xu C, He J, Chen Y, Cheng Y, Yang Q, Qiu H (2022) Graph-based convolution feature aggregation for retinal vessel segmentation. Simulation Modelling Practice and Theory 121:102653
31. Amin A, Shiu R (2001) Page segmentation and classification utilizing bottom-up approach. International Journal of Image and Graphics 1(02):345–361
32. Ha, J, Haralick, R.M, Phillips, IT: Document page decomposition by the bounding-box project. In: Proceedings of 3rd International Conference on Document Analysis and Recognition, vol 2, pp 1119–1122 (1995). IEEE
33. Ha, J, Haralick, R.M, Phillips, IT: Recursive x-y cut using bounding boxes of connected components. In: Proceedings of 3rd international conference on document analysis and recognition, vol 2, pp 952–955 (1995). IEEE
34. Lebourgeois, F, Bublinski, Z, Emptoz, H: A fast and efficient method for extracting text paragraphs and graphics from unconstrained documents. In: 11th IAPR international conference on pattern recognition. vol ii. conference b: pattern recognition methodology and systems, vol 1, pp 272–273 (1992). IEEE Computer Society
35. Shilman, M, Liang, P, Viola, P: Learning nongenerative grammatical models for document analysis. In: Tenth IEEE international conference on computer vision (ICCV'05) Volume 1, vol 2, pp 962–969 (2005). IEEE
36. Simon A, Pret J-C, Johnson AP (1997) A fast algorithm for bottom-up document layout analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(3):273–277
37. Mao, S, Rosenfeld, A, Kanungo, T: Document structure analysis algorithms: a literature survey. In: Document recognition and retrieval X, vol 5010, pp 197–207 (2003). International Society for Optics and Photonics
38. Wei, H, Baechler, M, Slimane, F, Ingold, R: Evaluation of svm, mlp and gmm classifiers for layout analysis of historical documents. In: 2013 12th international conference on document analysis and recognition, pp 1220–1224 (2013). IEEE
39. Schreiber, S, Agne, S, Wolf, I, Dengel, A, Ahmed, S: 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, pp 1162–1167 (2017). IEEE
40. He, D, Cohen, S, Price, B, Kifer, D, Giles, CL: Multi-scale multitask fcn for semantic page segmentation and table detection. In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR), vol 1, pp 254–261 (2017). IEEE
41. Siegel, N, Lourie, N, Power, R, Ammar, W: Extracting scientific figures with distantly supervised neural networks. In: Proceedings of the 18th ACM/IEEE on joint conference on digital libraries, pp 223–232 (2018)
42. Lee, J, Hayashi, H, Ohyama, W, Uchida, S: Page segmentation using a convolutional neural network with trainable co-occurrence features. In: 2019 International conference on document analysis and recognition (ICDAR), pp 1023–1028 (2019). IEEE
43. Agarwal, M, Mondal, A, Jawahar, C: Cdec-net: Composite deformable cascade network for table detection in document images. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp 9491–9498 (2021). IEEE
44. Paliwal, SS, Vishwanath, D, Rahul, R, Sharma, M, Vig, L: Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp 128–133 (2019). IEEE
45. Appalaraju, S, Jasani, B, Kota, B.U, Xie, Y, Manmatha, R: Docformer: End-to-end transformer for document understanding. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 993–1003 (2021)
46. Li, P, Gu, J, Kuen, J, Morariu, V.I, Zhao, H, Jain, R, Manjunatha, V, Liu, H: Selfdoc: Self-supervised document representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5652–5660 (2021)
47. Gu, Z, Meng, C, Wang, K, Lan, J, Wang, W, Gu, M, Zhang, L: Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4583–4592 (2022)
48. Jiao J, Tu W-C, Liu D, He S, Lau RW, Huang TS (2020) Formnet: Formatted learning for image restoration. IEEE Transactions on Image Processing 29:6302–6314
49. Huang, Y, Lv, T, Cui, L, Lu, Y, Wei, F: Layoutlmv3: Pre-training for document ai with unified text and image masking, pp 4083–4091 (2022). https://doi.org/10.1145/3503161.3548112
50. Lin, T-Y, Dollár, P, Girshick, R, He, K, Hariharan, B, Belongie, S: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2117–2125 (2017)
51. Hu, J, Shen, L, Sun, G: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141 (2018)
52. Dai, J, Qi, H, Xiong, Y, Li, Y, Zhang, G, Hu, H, Wei, Y: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 764–773 (2017)
53. Common-Object-in-Context: COCO Detection Evaluation. Website. https://cocodataset.org/#detection-eval (2021)
54. Jimeno Yepes, A, Zhong, P, Burdick, D: Icdar 2021 competition on scientific literature parsing. In: International conference on document analysis and recognition, pp 605–617 (2021). Springer
55. Zhang, S, Chi, C, Yao, Y, Lei, Z, Li, SZ: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9759–9768 (2020)
56. Li, J, Xu, Y, Lv, T, Cui, L, Zhang, C, Wei, F: Dit: Self-supervised pre-training for document image transformer, pp 3530–3539 (2022). https://doi.org/10.1145/3503161.3547911
57. Li, K, Wigington, C, Tensmeyer, C, Zhao, H, Barmpalios, N, Morariu, VI, Manjunatha, V, Sun, T, Fu, Y: Cross-domain document object detection: Benchmark suite and method. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12915–12924 (2020)
58. Li, M, Xu, Y, Cui, L, Huang, S, Wei, F, Li, Z, Zhou, M (2020) DocBank: A benchmark dataset for document layout analysis, pp 949–960. https://doi.org/10.18653/v1/2020.coling-main.82

59. Zhong, X, Tang, J, Yepes, AJ (2019) Publaynet: largest dataset ever for document layout analysis. In: 2019 International conference on document analysis and recognition (ICDAR), pp 1015–1022. IEEE