SRRV: A Novel Document Object Detector Based on Spatial-Related Relation and Vision

Hengyue Bi, Canhui Xu, Cao Shi, Guozhu Liu, Yuteng Li, Honghong Zhang, Jing Qu

***Abstract*—Document object detection is a challenging task due to layout complexity and object diversity. Most of existing methods mainly focus on vision information, neglecting representative inherent spatial-related relationship among document objects. To capture structural information and contextual dependencies, we propose a novel document object detector based on spatial-related relation and vision (SRRV). It consists of three parts: vision feature extraction network, relation feature aggregation network and result refinement network. Vision feature extraction network enhances information propagation of hierarchical feature pyramid by adopting feature augmentation paths. Then, relation feature aggregation network combines graph construction module and graph learning module. Specifically, graph construction module calculates spatial information from geometric attributes of region proposals to encode relation information, while graph learning module stacks Graph Convolutional Network (GCN) layers to aggregate relation information at global scale. Both the vision and relation features are fed into result refinement network for feature fusion and relational reasoning. Experiments on the PubLayNet, POD and Article Regions datasets demonstrate that spatial relation information improves the performance with better accuracy and more precise bounding box prediction.**

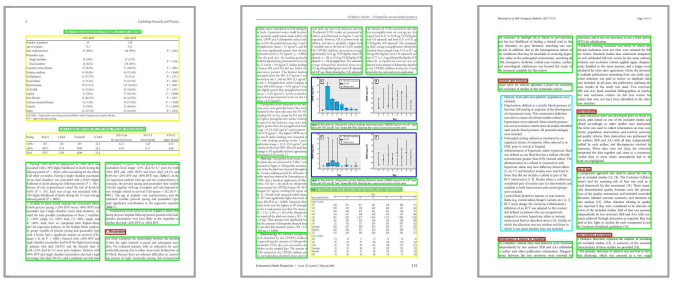
***Index Terms*—Document object detection, spatial-related relation, Graph Convolutional Network, feature representation, document layout analysis.**

I. Introduction

DOCUMENT image understanding involves document component detection and logical structure recovery in various levels such as character-level, line-level and block-level. Document object detection is to locate the page objects, such as text or non-text regions, which provides foundation for document image understanding. It could be widely applied in a variety of applications, such as information retrieval, document editing, text line transcription, document structure analysis.

Due to impressive feature extraction power, deep learning network has achieved significant progress in various computer vision tasks, such as image recognition [1], [2], semantic segmentation [3], [4], object detection [5], [6], salient object detection [7]–[9] and video saliency detection [10]–[12]. Recent advances in object detection, such as [13]–[17], have accelerated the progress of document object detection.

Mainstream deep learning based methods designed for natural scene images are adapted to explore the intrinsic characteristics of document images [18]–[21]. Besides these computer vision (CV) based document detection methods, some works put emphasis on combining CV based methods and natural language processing (NLP) based methods, which are known as multimodal networks [22], [23]. Yang et al. [22] present an end-to-end multimodal fully convolutional network to extract document semantic structure based on visual features and textual features with text embedding maps. VSR [23], as the top competitor of ICDAR Scientific Literature Parsing Competition (ICDAR-SLP) [24] in Document Layout Recognition challenge, presents a unified multimodal model to detect document objects by integrating both CV-based and NLP-based branch. In VSR, the NLP-based branch generates semantic features by parsing documents in PDFs. The CV-based branch extracts vision features by processing input document images. Multimodal methods might limit their application due to expensive computational cost and lack of multi-source information.



(a) (b) (c)

Fig. 1. Examples of document object detection task. (red: title, green: text, blue: figure, yellow: table, cyan: list)

Generally, document objects tend to have spatial-related relation and contextual dependencies. In Figure 1, we can observe obvious paired dependencies among logical labeled regions, such as table and table caption, figure and figure caption. Successive list items have numbered or bulleted marks and clear indentation in spatial layout. Besides, document page objects present inherent structural reading order. The above relational information works complementarily for human reading process, which inspires us to explore contextual information and inherent spatial relationship for boosting unimodal document object detection.

Recent unimodal works have attempted to apply relation information to document object detection. Li et al. [25] propose a hybrid method combining CNN and conditional random fields (CRF) [26] for capturing local context. The CRF model is used for supervised clustering with unary

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. The first two authors contributed equally to this work. Corresponding author: Guozhu Liu. (email: lgz_0228@163.com)

and binary potentials. With impressive learning power, Graph Convolutional Network [27], as an extension of CNN, could aggregate information from graph nodes and their neighbors, and achieve global information propagation.

In this paper, we propose a novel document object detection method based on spatial-related relation and vision (SRRV) to explore strong dependencies among objects. Particularly, SRRV consists of three sub-networks: vision feature extraction network (VFEN), relation feature aggregation network (RFAN) and result refinement network (RRN). Firstly, VFEN is to generate proposal candidates and their regional visual feature representation. Then, RFAN is to learn contextual information and inherent structural information via stacking GCN layers. Finally, to optimize the usage of VFEN and RFAN, RRN is to better fuse the obtained feature representation and achieve relational reasoning to bring contributions to precise detection results.

The main contributions of this work are summarized as follows:

- SRRV combines vision features and relation features to improve prediction accuracy and learning efficiency. Especially, RFAN integrates a graph construction module encoding relevant spatial regions and a graph learning module aggregating spatial information among document objects, which then can be propagated at global scale.
- To optimize the usage of visual and relation features, we design a result refinement network (RRN) to fuse features from two different distributions and gain additional inference.
- We evaluate the proposed method on three public available datasets. Extensive experimental results reveal that our proposed method can effectively boost document object detection performance.

II. Related work

A. Document Object Detection

Automatic document object detection plays an essential role in document image understanding and remains an open problem in image processing and computer vision. It aims to detect various document components which are of great importance for numerous application scenes. Early researches focus heavily on heuristic rules and handcrafted features, which are difficult to be used on document images with complex layouts and object contents [28]. In recent decade, due to its comprehensive feature extraction ability, deep learning based methods have been introduced to document object detection area. In ICDAR Page Object Detection Competition (ICDAR-POD) [29], most competitors applied CNN based methods and its variances to detect page objects with three categories including formula, figure and table.

In ICDAR-SLP Competition, document layout recognition is designated as task A which aims to promote more in-depth discussion and research. Some submitted methods try to introduce multimodal frameworks [24]. Team Hikvision Research Institute proposes VSR integrating both visual information and natural language model, which makes this team surpass all participants and rank first. Another team, Tomorrow Advancing Life utilizes Hybrid Task Cascade for Instance Segmentation method as baseline model and introduces LayoutLM [30] to optimize each text line. Meanwhile, team Simo adopts multimodal PDFMiner [31] to extract line coordinates of 'text' and 'title' for layout prediction refinement.

B. Relation Modeling

It is natural for humans to distinguish objects by using relation information which inspires researchers to explore object relation and gain additional inference. For example, Deng et al. [32] provide a unified framework to improve object classification by constructing a relation graph between labels. Li et al. [33] adopt knowledge graphs to describe relationship between multiple labels. Chen et al. [34] design an iterative reasoning framework to capture both spatial and semantic relationship between objects.

However, the above-mentioned methods rely on external handcraft knowledge graphs which require laborious preprocessing work. Recently, GCN [27] shows impressive learning power on graph-structured data in various tasks, such as [35]–[37]. Xu et al. [35] integrate a graph learner module to encode regional visual features by non-linear transformation and a spatial graph reasoning module with learnable spatial Gaussian kernels. Li et al. [36] introduce semantic label co-occurrence matrices to build edges between nodes which represent proposals in a heterogeneous graph. Chen et al. [37] adopt k Nearest Neighbors (kNN) to construct a local region for a point set, and use subtractions of central point and its neighbors to express their geometric relationship.

For document image processing, Liu et al. [38] apply GCN to compute visual embeddings from text segments generated by in-house Optical Character Recognition (OCR) system. Zhang et al. [39] utilize geometric attributes to construct local graphs which establish linkages between different text objects.

Inspired by the above-mentioned works, our goal is to learn a spatial awareness graph which embeds contextual information and inherent structural information to perform relation aggregation via GCN at global scale, and achieve relational reasoning among document objects.

III. Our approach

An overview of our proposed SRRV is illustrated in Figure 2. SRRV consists of three subnetworks, including vision feature extraction network (VFEN), relation feature aggregation network (RFAN) and result refinement network (RRN). In Figure 2(a), vision feature extraction network (VFEN) extracts augmented feature pyramid maps from backbone network, and generates candidate bounding boxes as region proposals. In Figure 2(b), relation feature aggregation network (RFAN) includes graph construction module and graph learning module. Graph construction module is to embed contextual spatial and inherent structural information into conventional visual convolutional networks. Then, graph learning module aggregates the relation information on the constructed graph, which is propagated globally. In Figure 2(c), result refinement network (RRN) integrates vision features and relation features, and it is capable of providing relational reasoning for ambiguous detection results.

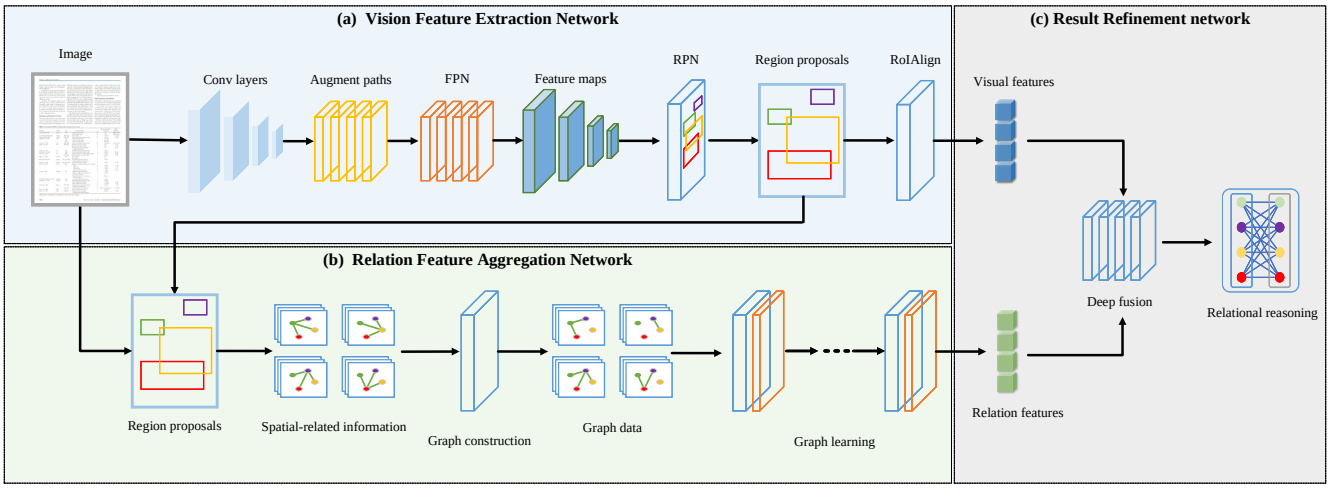


Fig. 2. The overall architecture of SRRV. (a) Vision feature extraction network (VFEN) utilizes backbone network to extract augmented feature maps from the input image, and then it produces region proposals. (b) Relation feature aggregation network (RFAN) includes graph construction module and graph learning module. Graph construction module is applied to build a relation graph by capturing spatial-related information among region proposals. Subsequently, graph learning module promotes information interaction based on the constructed graph, and updates graph embeddings to learn evolved feature representation for region proposals. (c) Result refinement network (RRN) is to integrate the obtained feature representation, and achieve relational reasoning.

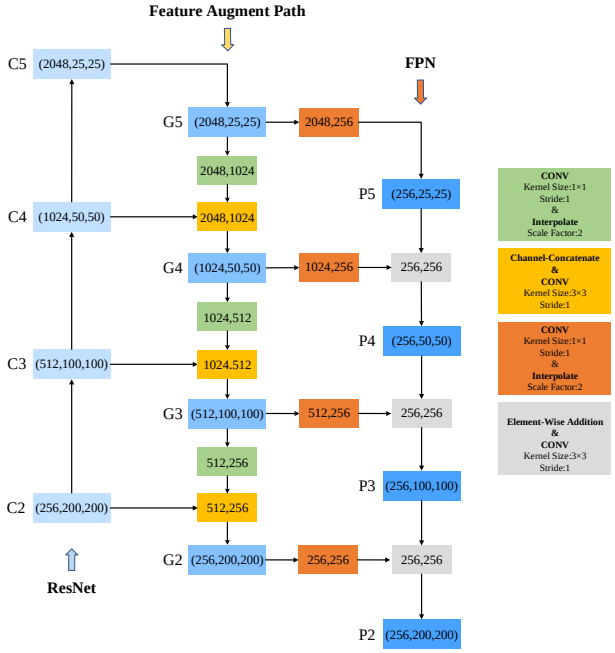


Fig. 3. Structural details of backbone network.

A. Vision Feature Extraction Network

To extract powerful visual features from both low-level and high-level layers, we propose feature augment paths to fuse multi-level features, as shown in Figure 3.

ResNet [40] is adopted as our basic backbone network. Feature maps generated by ResNet are denoted as $C = \{C_k \mid k \in \{1, , N\}\}$. Subsequently, to obtain informative feature maps, feature augment paths are extended to enable information propagation between feature pyramids. It starts from the highest level C_N and gradually fuses hierarchical information by a top-down path. The output feature maps $G = \{G_k \mid k \in \{1, , N\}\}$ are formulated in (1).

$$G_k = \begin{cases} C_N, & \text{if } k = N \\ Conv_2(Cat\left(Conv_1\left(G_{k+1}\right), C_k\right)), & \text{otherwise,} \end{cases} \tag{1}$$

where $Conv_1$ and $Conv_2$ are convolutional layers, Cat is the channel-concatenate operation. Feature Pyramid Network (FPN) [41] takes $G = \{G_k \mid k \in \{1, , N\}\}$ as input, and outputs semantically stronger feature maps $P = \{P_k \mid k \in \{1, , N\}\}$.

Region Proposal Network (RPN) is applied for region proposal generation. Anchor aspect ratios and strides are adapted to process document objects with various sizes and scales for generating candidate bounding boxes, which are formulated as (2).

$$n_i = (x_i, y_i, w_i, h_i)\,, \tag{2}$$

where each region proposal location is defined as (x_i, y_i, w_i, h_i). Region of Interest Align (RoIAlign) is applied to extract the visual feature representation f from the detected candidate bounding boxes.

B. Relation Feature Aggregation Network

An overview of our relation feature aggregation network (RFAN) is illustrated in Figure 4. Specifically, given an undirected graph $G = (V, E)$, graph node $v_i \in V$ is associated with initial features, and graph edge $e_{ij} \in E$ shows relation between pair-wise nodes (v_i, v_j). Graph construction module captures relative spatial position and geometric appearance similarity calculated from region proposals. Graph learning module learns to update node features through aggregating spatial-related information from its neighbors and propagating at global scale on the constructed graph.

1) Graph Construction Module: Given candidate bounding boxes, a graph is constructed to capture potential spatial and structure related information existing in document layout. As

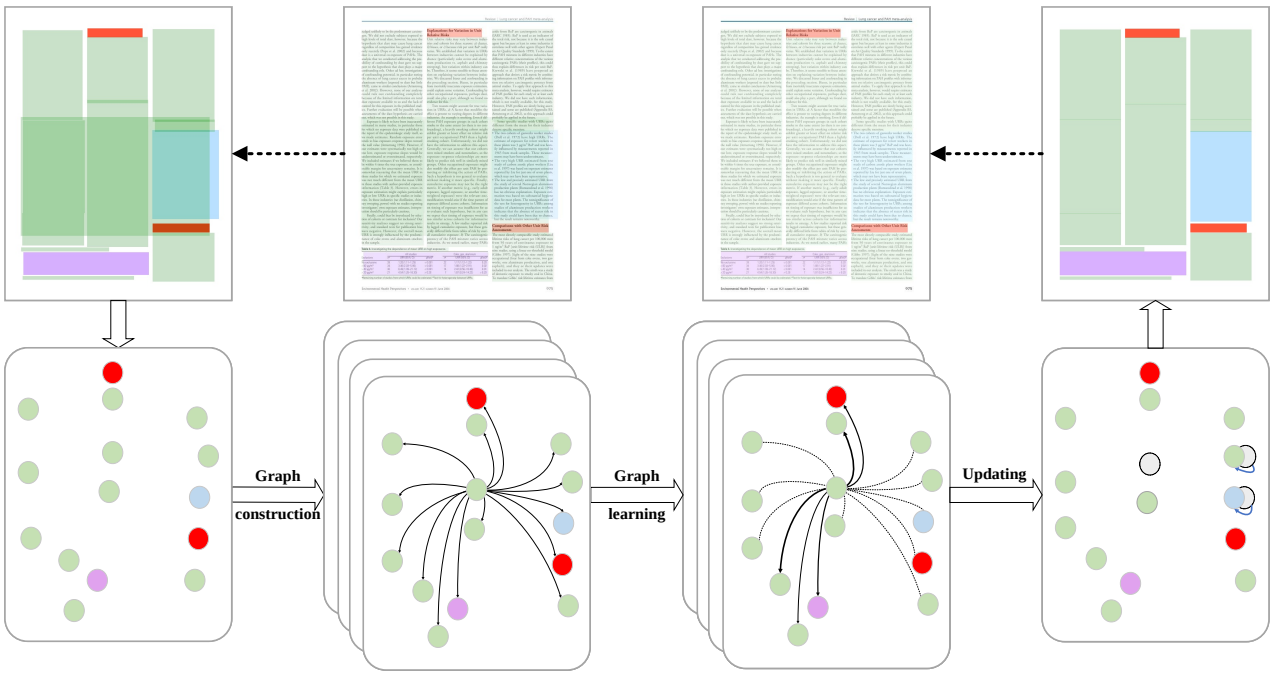


Fig. 4. The structure of relation feature aggregation network (RFAN). Set region proposals as graph nodes. Connected edges reflect adjacent relation between pair-wise graph nodes. Graph learning is to aggregate and propagate relation information on the constructed graph. Node feature representation could be updated through recursively passing messages from its neighbors. (red: title, green: text, brown: figure, purple: table, blue: list)

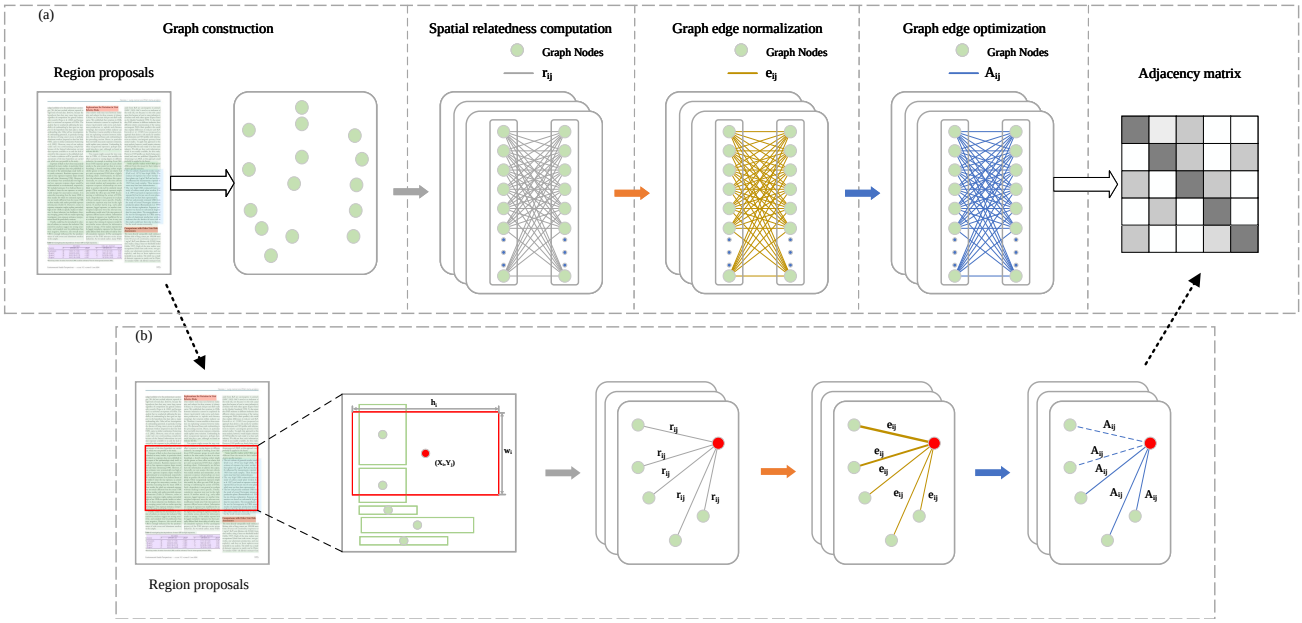


Fig. 5. (a) The flowchart of Graph Construction Module. Spatial relatedness r_{ij} establishes linkages between region proposals. Then, graph edges e_{ij} are calculated as the probabilities of pair-wise proposal linkages by normalizing the spatial relatedness r_{ij}. In addition, the constraints imported in Equation (8) are expected to retain more relevant relationship among proposals and prune noise linkages. Thus, we obtain adjacency matrix $A \in \mathbb{R}^{N \times N}$ of graph G. (b) A detailed example of the constructed graph.

illustrated in Figure 5(a), the graph construction is implemented among region proposals. For each region proposal i generated by RPN, the center coordinate $coord_i(x, y)$ is calculated using (3).

$$coord_i(x, y) = \left(x_i + \frac{w_i}{2}, y_i + \frac{h_i}{2} \right), \tag{3}$$

where bounding box (x_i, y_i, w_i, h_i) corresponds to the proposal i. Subsequently, spatial relatedness r_{ij} is encoded by measuring the similarity of relative position and shape factor between pair-wise graph nodes (v_i, v_j). This can be formulated as (4)-(6).

$$r_{ij} = \sqrt{\Delta coord_{ij} x^2 + \Delta coord_{ij} y^2}, \tag{4}$$

$$\Delta coord_{ij}x = \left(\frac{coord_i x}{w_i} - \frac{coord_j x}{w_j} \right), \tag{5}$$

$$\Delta coord_{ij}y = \left(\frac{coord_i y}{h_i} - \frac{coord_j y}{h_j} \right). \tag{6}$$

Then, graph edge e_{ij} is generated by applying normalization to relatedness r_{ij} with range of 0 to 1, as in (7).

$$e_{ij} = \exp(\frac{r_{ij}}{-\varepsilon \cdot \max_{k \in N} r_{ik}}), \tag{7}$$

where N is the number of region proposals, and ε is a modulating factor. Considering document page layout properties, constraints are imposed on the overlapped proposals, which tend to retain more informative edges and prune noise edges. It is computed as (8).

$$A_{ij} = \delta(i,j) \cdot e_{ij}, \tag{8}$$

where $A \in \mathbb{R}^{N \times N}$ is the adjacency matrix, $\delta(i,j)$ is an indicator function related to IoU threshold that equals 0 if proposal i and proposal j are overlapped with each other, and otherwise equals to 1. A constructed graph is shown in Figure 5(b).

2) Graph Learning Module: To tackle the graph learning problem, GCN, as an extension of CNN, is to operate convolutional in spectral or spatial domain. Herein, our graph learning module uses GCN for modeling relation features and information interaction to update node features. Based on the constructed graph, GCN takes the initial node features $X \in \mathbb{R}^{N \times D}$ and the adjacency matrix $A \in \mathbb{R}^{N \times N}$ as its inputs to aggregate relation features with spatial awareness. To be specific, each GCN layer is to learn a function g on the constructed graph G, as in (9)-(10):

$$H^l = g(X, A), \tag{9}$$

$$H^{l+1} = g(H^l, A), \tag{10}$$

where H^l is the input and H^{l+1} is the output of the l layer. After employing the convolutional operation, g can be written as (11)-(12):

$$H^{l+1} = \sigma\left(LH^{(l)}W^l\right), \tag{11}$$

$$L = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}, \tag{12}$$

where σ presents the $LeakyReLU$ activation function, W^l is the learned weight transformation matrix, $L \in \mathbb{R}^{N \times N}$is the normalized Laplace, $\tilde{A} = A + I_N$ is the adjacency matrix of the constructed graph G with added self-connections, $\tilde{D} \in \mathbb{R}^{N \times N}$ is the degree matrix.

C. Result Refinement Network

Result refinement network (RRN) consists of two modules including deep fusion module (DFM) and relational reasoning module (RRM), which will be elaborated.

1) Deep Fusion Module: In [42], it claims that different types of features have different value ranges and distributions, which motivates us to develop a suitable fusion strategy to integrate vision features and relation features. Given the regional visual features f and the relation features H^{l+1}:

$$\tilde{f} = fc(Cat(H^{1+1}, f)), \tag{13}$$

where $\tilde{f}$ is the fused features, Cat is the channel-concatenate operation, and fc denotes the fully-connected layers liked Mask R-CNN [17].

2) Relational Reasoning Module: RRM aims to select the highly spatial-related region proposals. A multi-layer perceptron (MLP) is utilized for encoding the spatial-related relationship, which generates a score matrix corresponding to the region proposals. The computation process can be written as (14):

$$H^{l+1} \overset{Linear}{\longrightarrow} (N, d_1) \overset{Linear}{\longrightarrow} (N, d_2) \overset{\alpha}{\longrightarrow} (N, d_3) \longrightarrow S, \tag{14}$$

where $Linear$ denotes the linear regression operation, α is the $sigmoid$ function computing the spatial-related probabilities, S is the score matrix, d_1, d_2 and d_3 are the output channels. For the obtained matrix S, a scale factor ρ is to select proposals with higher spatial-related probabilities. Refined region proposals are thus provided for classification and box regression branches.

IV. Experiments

In this section, a series of experiments are conducted to validate the effectiveness of SRRV on three widely used datasets including POD, Article Regions and PubLayNet.

A. Datasets and Evaluation Metrics

Three public available and widely used datasets including POD [29], Article Regions [47] and PubLayNet [48] are described in detail. PubLayNet is a large-scale dataset for document layout analysis. It consists of 360K document images with 5 categories (text, title, list, table, figure), including 335,703 training images, 11,245 validation images, 20 mini validation images and 11,405 test images. It adopts the standard mean average precision (mAP) @ IoU $[0.5, 0.95]$ as evaluation metric. Article Regions [47] consists of 822 document images with 9 categories (title, authors, abstract, body, figure, figure caption, table, table caption and reference), and the whole dataset is split into 600 and 222 images for training and validation. It uses the mAP at IoU threshold 0.5 as evaluation metric. POD [29], as the ICDAR POD competition dataset, consists of 2,417 document images with 3 categories (Formula, Table, Figure), in which 1,600 images are used as training set and 817 images are used as test set. It adopts the mAP at two IoU thresholds (0.6 and 0.8) as evaluation metrics.

B. Implementation Details

We implement our proposed SRRV with Pytorch framework. Specifically, three different backbone networks are extended in SRRV, including ResNet-50, ResNet-101, and ResNeXt-101-32x8d [49] that are all pre-trained on ImageNet

TABLE I
PERFORMANCE COMPARISON OF THE PROPOSED SRRV AND THREE DOCUMENT OBJECT DETECTION METHODS ON PUBLAYNET VALIDATION DATASET.

Method	Backbone	mAP	text	title	list	table	figure
CDDOD [43]	ResNeXt-32x8d-101	0.922	0.923	0.913	0.913	0.934	0.927
CDeCNet [44]	ResNeXt-32x8d-101	0.905	0.915	0.840	0.895	0.969	0.906
VSR [23]	ResNeXt	**0.957**	**0.967**	**0.931**	0.947	0.974	0.964
SRRV	ResNeXt-32x8d-101	0.950	0.958	0.901	**0.950**	**0.976**	**0.967**

TABLE II
PERFORMANCE COMPARISON OF THE PROPOSED SRRV AND FOUR COMMON OBJECT DETECTION METHODS ON PUBLAYNET VALIDATION DATASET.

Method	Backbone	mAP	APs	APm	APl	text	title	list	table	figure
Mask R-CNN	ResNet-50	0.888	0.374	0.675	0.921	0.908	0.801	0.860	0.950	0.921
ATSS	ResNet-50	0.853	0.280	0.617	0.896	0.889	0.754	0.816	0.931	0.876
Cascade R-CNN	ResNet-50	0.908	0.365	**0.717**	**0.948**	0.909	0.835	**0.886**	**0.969**	0.939
Faster R-CNN(Baseline)	ResNet-50	0.867	0.364	0.663	0.901	0.895	0.801	0.820	0.920	0.901
SRRV	ResNet-50	**0.909**	**0.395**	0.686	0.935	**0.919**	**0.840**	0.876	0.963	**0.947**
Mask R-CNN	ResNet-101	0.899	0.362	0.729	0.937	0.908	0.830	0.854	0.960	0.941
ATSS	ResNet-101	0.865	0.314	0.613	0.905	0.892	0.764	0.840	0.934	0.893
Cascade R-CNN	ResNet-101	0.914	0.390	0.745	**0.951**	0.922	0.845	0.886	0.967	0.949
Faster R-CNN(Baseline)	ResNet-101	0.880	0.363	0.726	0.913	0.903	0.812	0.834	0.941	0.912
SRRV	ResNet-101	**0.919**	**0.423**	**0.762**	0.940	**0.926**	**0.849**	**0.900**	**0.968**	**0.953**
Mask R-CNN	ResNeXt-32x8d-101	0.932	0.456	0.819	0.967	0.930	0.860	0.935	0.973	0.964
ATSS	ResNeXt-32x8d-101	0.928	0.406	0.763	0.964	0.937	0.837	0.936	0.974	0.954
Cascade R-CNN	ResNeXt-32x8d-101	0.935	0.432	0.829	0.964	0.943	0.875	0.924	0.972	0.935
Faster R-CNN(Baseline)	ResNeXt-32x8d-101	0.928	0.457	0.832	0.965	0.929	0.852	0.932	0.965	0.963
SRRV	ResNeXt-32x8d-101	**0.950**	**0.470**	**0.842**	**0.972**	**0.958**	**0.901**	**0.950**	**0.976**	**0.967**

[50] to initialize parameters. For RPN, we design 5 scales and 7 aspect ratios for anchor generation. Each mini-batch is set to 2, so each GPU has two images and each image has 512 sampled RoIs with a positive-negative ratio of 1:3 [51]. We train our SRRV on one Nvidia GTX 2080Ti GPU using stochastic gradient descent (SGD) with a momentum of 0.9 and weight decay of 0.0001. For backbone ResNet-50, we train our model for 75k iterations, with an initial learning rate of 0.0025 which is reduced by a factor of 10 at 40k iterations and 60k iterations. For backbone ResNet-101, we train our model for 75.5k iterations, with an initial learning rate of 0.0025 which is reduced by a factor of 10 at 40k iterations and 60k iterations. For backbone ResNeXt-101-32$\times$8d, we train our model for 270k iterations, with an initial learning rate of 0.0009 which is reduced by a factor of 10 at 210k iterations and 250k iterations. Specifically, the training model of Article Regions lasts for 10K iterations using ResNet-101. All of the experiments are conducted on an unbuntu20.04 workstation with an Intel(R) Xeon(R) Silver 4210 2.20GHz CPU 64GB RAM.

C. Results on PubLayNet

PubLayNet is a large-scale dataset consisting of 360K document images. The ICDAR 2021 competition SLP task A uses PubLayNet as its competition dataset. VSR [23], as the winner of this competition, is compared with SRRV in Table I. VSR [23] integrates both CV based and NLP based methods, while our proposed SRRV explores the potential of CV based method itself. In Table I, SRRV achieves 0.950 mAP on PubLayNet validation dataset, which is higher than CDDOD [43] and CDeCNet [44], and is 0.7% lower than VSR [23]. But SRRV is superior to VSR [23] on the list, table and figure categories. While CDDOD [43] and CDeCNet [44] only focus on extracting vision features, SRRV performs better by exploring relation features calculated from spatial relatedness among objects which are further integrated with vision features. In Table I, it can be observed that SRRV shows competitive performance on table, figure and list, even without utilizing multimodal information or additional natural language information. This indicates that vision information based deep learning network still has the potential to be optimized.

Table II shows the performance comparisons of SRRV and four common object detection methods on different backbones, including Faster R-CNN [13], Mask R-CNN [17], ATSS [39] and Cascade R-CNN [52]. When using ResNet-50 and ResNet-101 as the basic backbone network, SRRV surpasses the above-mentioned methods. And the performances of SRRV are improved by 4.2% and 3.9% on mAP compared to the baseline model, respectively. When using ResNeXt, SRRV outperforms all the compared methods, by around average 2% higher on mAP. With three different mainstream backbone networks, SRRV is able to gain performance improvement steadily.

D. Results on POD

The proposed SRRV is compared with thirteen document object detection methods in Table III on POD test dataset to comprehensively analyze the effectiveness of our proposed

TABLE III
PERFORMANCE COMPARISON OF THE PROPOSED SRRV AND THIRTEEN DOCUMENT OBJECT DETECTION METHODS ON POD TEST DATASET.

Method	AP(IoU=0.6)				AP(IoU=0.8)			
	Formula	Table	Figure	mAP	Formula	Table	Figure	mAP
NLPR-PAL	0.839	0.933	0.849	0.874	0.816	0.911	0.805	0.844
icstpku	0.849	0.753	0.679	0.760	0.815	0.697	0.597	0.703
FastDetectors	0.474	0.925	0.392	0.597	0.427	0.884	0.365	0.559
Vislnt	0.524	0.914	0.781	0.740	0.117	0.795	0.565	0.492
SOS	0.537	0.931	0.785	0.751	0.109	0.737	0.518	0.455
UITVN	0.193	0.924	0.786	0.634	0.061	0.695	0.554	0.437
Matiai-ee	0.116	0.781	0.325	0.407	0.005	0.626	0.134	0.255
HustVision	0.854	0.938	0.853	0.882	0.293	0.796	0.656	0.582
Li eta [25]	0.878	0.946	0.896	0.907	0.863	0.923	0.854	0.880
FFD [45]	0.897	/	0.886	0.892	0.776	/	0.794	0.785
CDDOD [43]	0.857	0.901	0.821	0.860	0.750	0.825	0.805	0.793
GOD [46]	0.901	0.942	0.841	0.895	0.832	0.924	0.813	0.856
CDeCNet [44]	0.821	0.922	0.884	0.875	0.803	0.915	0.832	0.850
SRRV	**0.962**	**0.963**	**0.912**	**0.946**	**0.951**	**0.953**	**0.865**	**0.923**

TABLE IV
PERFORMANCE COMPARISON OF THE PROPOSED SRRV AND THREE METHODS ON ARTICLE REGIONS VALIDATION DATASET.

Method	Backbone	title	author	abstract	body	figure	figure caption	table	table caption	reference	mAP
CDDOD [43]	ResNet-101	0.886	0.632	0.876	0.933	0.829	0.879	0.823	0.696	0.943	0.833
VSR [23]	ResNet-101	**1**	**0.940**	0.950	**0.991**	**0.953**	0.945	**0.961**	0.846	0.923	0.945
Faster R-CNN(Baseline)	ResNet-101	0.964	0.66	0.901	0.98	0.932	0.889	0.945	0.671	0.901	0.871
SRRV	ResNet-101	0.959	0.858	**0.992**	0.990	0.935	**0.970**	0.951	**0.878**	**0.999**	**0.948**

method. SRRV has achieved higher accuracy than the thirteen compared methods. Furthermore, when IoU threshold takes the value of 0.8, the performance on formulation category obtains a significant improvement, which is 8.8% higher than Li et al. [25]. This demonstrates that SRRV is capable of boosting the accuracy of small document object detection (such as formulation category in POD) by aggregating relation information at global scale.

E. Results on Article Regions

Article Regions consists of nine categories, where inter-classed relation is more complicated. In Table IV, SRRV outperforms all the compared methods on mAP. Especially on reference category, SRRV is nearly 10% higher than VSR [23]. It is attributed to the encoded spatial relation and relational reasoning. Compared with CDDOD [43], SRRV performs better by making use of the contextual spatial information among objects within a document page, such as the pair-wise figure and figure caption, table and table caption.

In Figure 6, qualitative examples of SRRV, Faster R-CNN [13] and CDDOD [43] are illustrated. In Figure 6 (c-3) and (c-4), the yellow table caption below the purple table is misclassified as figure caption. From Figure 6 (c-2), we can observe that our method SRRV detects these objects correctly. Furthermore, some imprecise boxes overlaid with each other can be spotted in Figure 6 (a-4) and (c-4). In this case, SRRV obtains higher precision of box regression with the contribution of refined proposals. In Figure 6 (b-3) and (b-4), a brown figure box is not detected by CDDOD [43] and Faster R-CNN, but SRRV is able to identify.

Fig. 6. Qualitative results of the proposed SRRV and the compared methods on Article Regions.

TABLE V
ABLATION STUDY BASED ON DIFFERENT SUBNETWORKS OF SRRV ON POD TEST DATASET.

VFEN	RFAN	RRN	Datasets	AP(IoU=0.6)				AP(IoU=0.8)			
				Formula	Table	Figure	mAP	Formula	Table	Figure	mAP
			POD	0.938	0.943	0.870	0.917	0.903	0.913	0.839	0.885
√			POD	0.942	0.946	0.911	0.933	0.918	0.914	0.846	0.893
	√		POD	0.958	0.952	0.909	0.940	0.942	0.935	0.846	0.908
√	√		POD	0.957	0.956	0.910	0.941	0.943	0.945	0.842	0.910
√	√	√	POD	**0.962**	**0.963**	**0.912**	**0.946**	**0.951**	**0.953**	**0.865**	**0.923**

TABLE VI
ABLATION STUDY BASED ON DIFFERENT SUBNETWORKS OF SRRV ON PUBLAYNET VALIDATION DATASET.

VFEN	RFAN	RRN	Datasets	mAP	APs	APm	APl
			PubLayNet	0.880	0.363	0.726	0.913
√			PubLayNet	0.901	0.378	0.748	0.946
	√		PubLayNet	0.913	0.403	0.753	0.953
√	√		PubLayNet	0.915	0.406	**0.803**	0.945
√	√	√	PubLayNet	**0.919**	**0.423**	0.762	**0.946**

TABLE VII
EFFECTS OF DIFFERENT FEATURE FUSION STRATEGIES ON ARTICLE REGIONS VALIDATION DATASET.

method	Datasets	mAP
element-wise concatenation	Article Regions	0.929
element-wise addition	Article Regions	0.927
DFM	Article Regions	**0.940**

F. Ablation Study

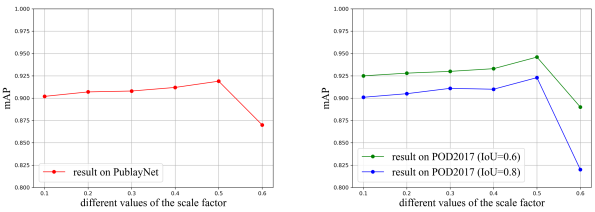
Herein, we conduct comprehensive ablation studies to investigate the performance of the subnetworks in SRRV. Table V and Table VI show the effectiveness of different subnetworks (VFEN, RFAN and RRN) on POD test and PubLayNet validation dataset, respectively.

1) Effectiveness of Vision Feature Extraction Network (VFEN): Compared with the baseline model Faster R-CNN, RFAN boosts the detection performance from 0.880 to 0.901 mAP on PubLayNet validation dataset. And it obtains similar improvements on POD test dataset. This could be attributed to that our proposed feature augment paths are able to advance feature extraction ability of basic backbone network.

2) Effectiveness of Relation Feature Aggregation Network (RFAN): RFAN is the most important subnetwork of our proposed SRRV. In Table V, when the IoU threshold takes the value of 0.8, RFAN achieves 0.908 mAP which is 2.3% higher than Faster R-CNN. In Table VI, results on PubLayNet show that RFAN achieves 0.913 mAP and 0.403 APs, which are respective 3.3% and 4% higher than the baseline model. The experimental results indicate the importance of encoding spatial relation among document objects and propagating relation information at global scale.

3) Effectiveness of Result Refinement Network (RRN): In this part, we mainly discuss how to leverage vision features and relation features. From Table V and Table VI, our model equipped with VFEN and RFAN could improve document object detection to some extent. But our full model (equipped with VFEN, RFAN and RRN) yields further improvement. RRN includes a deep fusion module (DFM) integrating features from the VFEN and RFAN subnetworks and a relational reasoning module (RRM) achieving relational reasoning. To assess the importance of DFM and RRM, we evaluate the effectiveness of different feature fusion strategies in DFM and different values of scale factor ρ in RRM.

a) Effectiveness of Deep Fusion Module (DFM): In Table VII, three types of concatenation methods are implemented. As can be seen, the performance of DFM is higher than the element-wise concatenation and element-wise addition with 0.11 and 0.13 mAP, respectively.



(a) Comparisons on PubLayNet. (b) Comparisons on POD.

Fig. 7. Accuracy comparisons with different values of ρ.

b) Effectiveness of Relational Reasoning Module (RRM): Set scale factor ρ ranging from 0.1 to 0.6. In Figure 7, We can see that when $\rho = 0.5$ SRRV achieves best performance on both PubLayNet and POD. This could be attributed to the fact that some redundant candidate bounding boxes of low spatial relatedness probability have been pruned. However, when the scale factor is too strict, the accuracy drops quickly since the target bounding boxes of high spatial relatedness probability could be filtered out.

Figure 8 depicts the qualitative comparisons of RFAN with the baseline model Faster R-CNN on PubLayNet validation dataset. Especially, the constructed graph structure from RFAN is visualized in Figure 8(a-2), (b-2), (c-2) and (d-2). The centroids of regions are connected by the calculated graph edges. Edge thickness corresponds to the scaled edge weights. In Figure 8(a-1), the text region in bottom-right of the page is undetected. RFAN could guide the detector by strong linkages between the text region and its neighbors. Thus, the undetected problem is alleviated. Among three blue list objects in Figure 8(b-1), two of them are misclassified by using the baseline

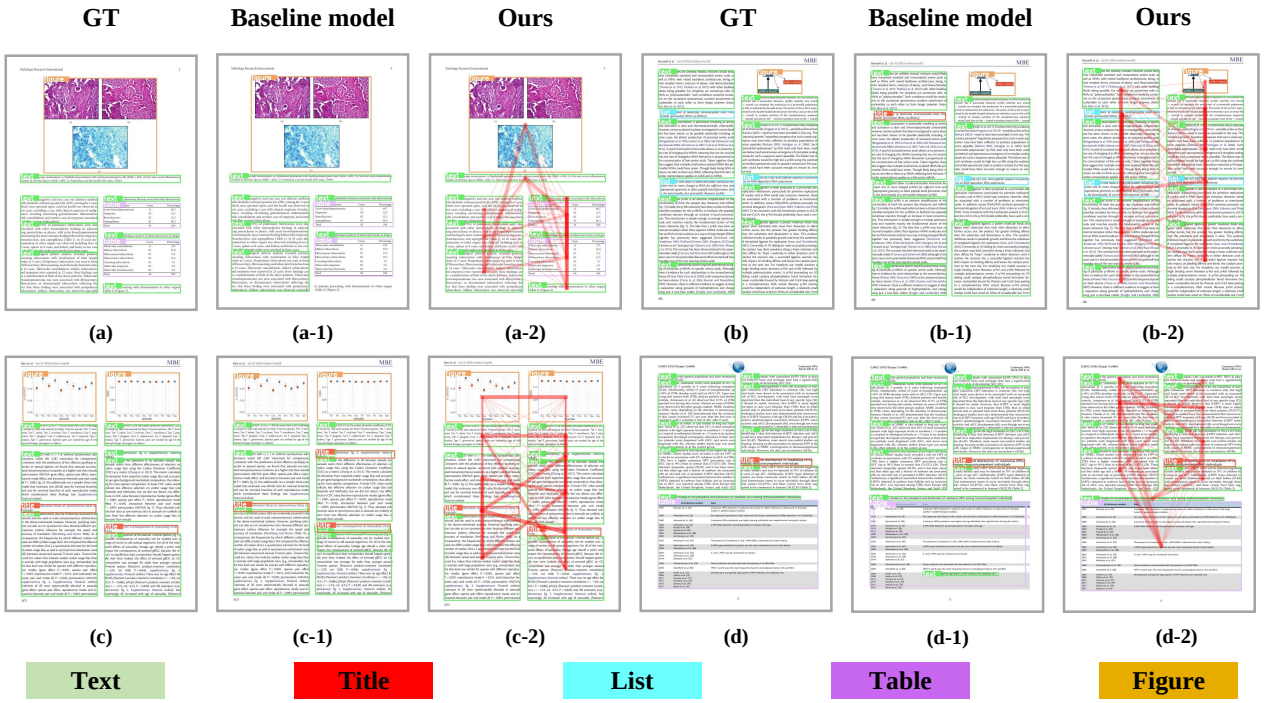
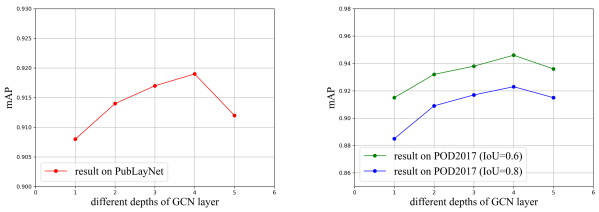


Fig. 8. Qualitative results of the proposed RFAN and the baseline model on PubLayNet. The centroids of regions are connected by the calculated graph edges. Edge thickness corresponds to the scaled edge weights.



(a) Comparisons on PubLayNet. (b) Comparisons on POD.

Fig. 9. Accuracy comparisons with different depths of GCN layer.

model without relation embedding. In Figure 8(b-2), these lists with high spatial similarities are connected obviously by the graph edges. After aggregating relation information, their feature representation thus is enhanced, which might enable RFAN to covert the imprecise prediction. In Figure 8(c-1) and (d-1), baseline Faster R-CNN suffers from overlap issue. RFAN enables the graph learning module to propagate the global spatial information among region proposals. Moreover, the imported constraints would retain informative edges and the detection results are less likely to have overlaps.

4) Effectiveness of Different Depths of GCN: An important reason behind the great success of CNN is the degradation problem has been solved, which makes deep CNN to be reliably trained. However, it is not suitable to stack more GCN layers due to over-smoothing problem [53]. To discuss the depth of GCN, an ablation study is given to analyze its effectiveness. In Figure 9, the performance increases to optimal point at 4 stacked layers. To further deepen GCN layers is likely to deteriorate performance.

G. Computation Efficiency Analysis

To evaluate the computation efficiency of SRRV, we investigate the average running-time of our method SRRV, CDDOD and CDeCNet on PubLayNet validation dataset. SRRV takes an average of 0.195 second to process one image with 1333×800 resolution, which is slower than the compared methods. But, SRRV gains better accuracy with reasonable sacrifice on speed.

To further explore computation time distribution, we calculate the consumption time of each designed subnetwork of SRRV, including vision feature extraction network (VFEN), relation feature aggregation network (RFAN) and result refinement network (RRN). RFAN seems to be a bottleneck since it consumes most of the computation time, which consists of graph construction module and graph learning module. SRRV aims to learn a dynamic learnable graph for each image which might inevitably bring additional computation cost. In the future, the graph construction could be accelerated by designing faster computation algorithms, and the graph learning should be optimized to reduce computational complexity.

V. Conclusion

A novel document object detection method SRRV is proposed to optimize the performance of unimodal visual convolutional network by embedding spatial relation information. To begin with, vision feature extraction network is to extract more informative vision features by feature augment paths. Then, relation feature aggregation network combines graph

construction module and graph learning module. Graph construction module calculates spatial information from geometric attributes to encode relationship. While graph learning module stacks GCN layers to aggregate relation information propagated among objects at global scale. At last, result refinement network includes deep fusion module integrating the features from two different distributions effectively and relational reasoning module for relational inference to bring contributions to filter the ambiguous results. Extensive experimental results demonstrate the importance of relation information extraction, which improves the document object detection accuracy and simultaneously benefits small object detection. In comparison with state-of-the-art methods, SRRV has more effective and robust performances on three widely used datasets.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[2] L. Herranz, S. Jiang, and R. Xu, "Modeling restaurant context for food recognition," *IEEE Transactions on Multimedia*, vol. 19, no. 2, pp. 430–440, 2017.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[4] B. Kang, Y. Lee, and T. Q. Nguyen, "Depth-adaptive deep neural network for semantic segmentation," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2478–2490, 2018.

[5] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[6] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, and S. Yan, "Attentive contexts for object detection," *IEEE Transactions on Multimedia*, vol. 19, no. 5, pp. 944–954, 2017.

[7] Y. Zhou, A. Mao, S. Huo, J. Lei, and S.-Y. Kung, "Salient object detection via fuzzy theory and object-level enhancement," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 74–85, 2019.

[8] G. Ma, C. Chen, S. Li, C. Peng, A. Hao, and H. Qin, "Salient object detection via multiple instance joint re-learning," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 324–336, 2020.

[9] C. Chen, J. Wei, C. Peng, and H. Qin, "Depth-quality-aware salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 2350–2363, 2021.

[10] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao, "Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3156–3170, 2017.

[11] C. Chen, S. Li, H. Qin, Z. Pan, and G. Yang, "Bilevel feature learning for video saliency detection," *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3324–3336, 2018.

[12] Y. Li, S. Li, C. Chen, A. Hao, and H. Qin, "Accurate and robust video saliency detection via self-paced diffusion," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1153–1167, 2019.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

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

[15] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[16] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2874–2883.

[17] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[18] X. Yi, L. Gao, Y. Liao, X. Zhang, R. Liu, and Z. Jiang, "Cnn based page object detection in document images," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, 2017, pp. 230–235.

[19] S. Schreiber, S. Agne, I. Wolf, A. Dengel, and S. Ahmed, "Deepdesrt: Deep learning for detection and structure recognition of tables in document images," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, 2017, pp. 1162–1167.

[20] C.-h. Xu, C. Shi, and Y.-n. Chen, "End-to-end dilated convolution network for document image semantic segmentation," *Journal of Central South University*, vol. 28, no. 6, pp. 1765–1774, 2021.

[21] C. Xu, C. Shi, H. Bi, C. Liu, Y. Yuan, H. Guo, and Y. Chen, "A page object detection method based on mask r-cnn," *IEEE Access*, vol. 9, pp. 143 448–143 457, 2021.

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

[23] P. Zhang, C. Li, L. Qiao, Z. Cheng, S. Pu, Y. Niu, and F. Wu, "Vsr: A unified framework for document layout analysis combining vision, semantics and relations," 2021.

[24] A. J. Yepes, X. Zhong, and D. Burdick, "Icdar 2021 competition on scientific literature parsing," 2021.

[25] X.-H. Li, F. Yin, and C.-L. Liu, "Page object detection from pdf document images by deep structured prediction and supervised clustering," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 3627–3632.

[26] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[27] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2017.

[28] X. Tao, Z. Tang, and C. Xu, "Contextual modeling for logical labeling of pdf documents," *Computers & Electrical Engineering*, vol. 40, no. 4, pp. 1363–1375, 2014.

[29] L. Gao, X. Yi, Z. Jiang, L. Hao, and Z. Tang, "Icdar2017 competition on page object detection," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, 2017, pp. 1417–1422.

[30] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "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*, 2020, pp. 1192–1200.

[31] Y. Shinyama, "Pdfminer: Python pdf parser and analyzer," *Retrieved on*, vol. 11, 2015.

[32] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam, "Large-scale object classification using label relation graphs," in *European conference on computer vision*. Springer, 2014, pp. 48–64.

[33] Q. Li, M. Qiao, W. Bian, and D. Tao, "Conditional graphical lasso for multi-label image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[34] X. Chen, L.-J. Li, L. Fei-Fei, and A. Gupta, "Iterative visual reasoning beyond convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7239–7248.

[35] H. Xu, C. Jiang, X. Liang, and Z. Li, "Spatial-aware graph relation network for large-scale object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9298–9307.

[36] Z. Li, X. Du, and Y. Cao, "Gar: Graph assisted reasoning for object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1295–1304.

[37] J. Chen, B. Lei, Q. Song, H. Ying, D. Z. Chen, and J. Wu, "A hierarchical graph network for 3d object detection on point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 392–401.

[38] X. Liu, F. Gao, Q. Zhang, and H. Zhao, "Graph convolution for multimodal information extraction from visually rich documents," *arXiv preprint arXiv:1903.11279*, 2019.

[39] S.-X. Zhang, X. Zhu, J.-B. Hou, C. Liu, C. Yang, H. Wang, and X.-C. Yin, "Deep relational reasoning graph network for arbitrary shape text detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[41] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.
[42] S. Jiang, W. Min, L. Liu, and Z. Luo, "Multi-scale multi-view deep feature aggregation for food recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 265–276, 2019.
[43] K. Li, C. Wigington, C. Tensmeyer, H. Zhao, N. Barmpalios, V. I. Morariu, V. Manjunatha, T. Sun, and Y. Fu, "Cross-domain document object detection: Benchmark suite and method," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 915–12 924.
[44] M. Agarwal, A. Mondal, and C. Jawahar, "Cdec-net: Composite deformable cascade network for table detection in document images," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 9491–9498.
[45] J. Younas, S. T. R. Rizvi, M. I. Malik, F. Shafait, P. Lukowicz, and S. Ahmed, "Ffd: Figure and formula detection from document images," in *2019 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2019, pp. 1–7.
[46] R. Saha, A. Mondal, and C. Jawahar, "Graphical object detection in document images," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 51–58.
[47] C. Soto and S. Yoo, "Visual detection with context for document layout analysis," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3464–3470.
[48] X. Zhong, J. Tang, and A. Jimeno Yepes, "Publaynet: Largest dataset ever for document layout analysis," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, pp. 1015–1022.
[49] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.
[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.
[51] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.
[52] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
[53] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Thirty-Second AAAI conference on artificial intelligence*, 2018.

Hengyue Bi is currently pursuing the M.S. degree with the School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, China. His research interests include computer vision and machine learning.

Canhui Xu received her Ph.D. degree from Central South University, China, in 2011. She is currently working in the School of Information Science and Technology, Qingdao University of Science and Technology, China. She was a postdoctoral research fellow at Peking University from 2012 to 2014. She was a visiting scholar at Arizona State University, USA, from 2019 to 2020, and a visiting Ph.D. student at Imperial Collage London, UK, from 2009 to 2010. Her research interests include deep learning, document layout analysis and image understanding.

Cao Shi received Ph. D degree in 2011 from Central South University, and now works with the School of Information Science and Technology, Qingdao University of Science and Technology, China. He was a postdoctoral research fellow at Peking University from 2011 to 2013. His research interests include image, video processing, and artificial intelligence.

Guozhu Liu received the Ph.D. degree in computer science and technology from the Qingdao University of Science and Technology, Qingdao, China. He is currently a Professor of computer science with the School of Information Science and Technology, Qingdao University of Science and Technology. His current research interests include deep learning, recommender system and intelligent software analysis.

Yuteng Li is currently pursuing the M.S. degree with the School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, China. His research interests include deep learning, computer vision and image processing.

Honghong Zhang is currently pursuing the M.S. degree with the School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, China. Her research interests include artificial intelligence, computer vision, image processing.

Jing Qu is currently pursuing the M.S. degree with the School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, China. Her research interests include pattern recognition and machine learning.