
PARAGRAPH2GRAPH: A GNN-BASED FRAMEWORK FOR LAYOUT PARAGRAPH ANALYSIS

Shu Wei

Datagrand Tech Inc.
weishucv@gmail.com

Nuo Xu

Datagrand Tech Inc.
nxu8@outlook.com

Deng Huang

Datagrand Tech Inc.
yangyu4608@gmail.com

Xiang Gao

Datagrand Tech Inc.
gaoxiang@datagrand.com

ABSTRACT

Document layout analysis has a wide range of requirements across various domains, languages, and business scenarios. However, most current state-of-the-art algorithms are language-dependent, with architectures that rely on transformer encoders or language-specific text encoders, such as BERT, for feature extraction. These approaches are limited in their ability to handle very long documents due to input sequence length constraints and are closely tied to language-specific tokenizers. Additionally, training a cross-language text encoder can be challenging due to the lack of labeled multilingual document datasets that consider privacy. Furthermore, some layout tasks require a clean separation between different layout components without overlap, which can be difficult for image segmentation-based algorithms to achieve. In this paper, we present Paragraph2Graph, a language-independent graph neural network (GNN)-based model that achieves competitive results on common document layout datasets while being adaptable to business scenarios with strict separation. With only 19.95 million parameters, our model is suitable for industrial applications, particularly in multi-language scenarios. We are releasing all of our code and pretrained models at this repo.

Keywords GNN · Language-independent · Document Layout · Layout Paragraph · Generalization

1 Introduction

Document layout analysis is an important task for very-rich document understanding. Given the availability to the text bounding boxes, text info and document image, most current works either integrate all modalities together with BERT-like encoders [1][2][3][4] or simply using visual information [5] [6] to model the task as an object detection problem. While effective, industrial applications need to consider very-long multilingual paragraphs, which a BERT-like encoder fails to hold due to the limitation of input sequence length and lack of multilingual document dataset. Moreover, some scenarios expecting a clear separation between layout components make image segmentation-based algorithms hard to adapt due to vague boundaries. Although post-processing can handle the problems, hand-craft rules make the pipeline complicated and hard to be maintained. In contrast, graph neural networks (GNNs) can offer a promising alternative approach that does not rely on language models.

With this work, we propose Paragraph2Graph, a language-independent GNN model to address these limitations. Fig.1 shows the overall architecture. We first encode image features with a pre-trained CNN backbone. Since each OCR box can be regarded as a spatially-separated node of a graph, we therefore incorporate the 2d OCR text coordinates, denoted as layout modality, and image features together as node features. Then, we build our neural network with DGCNN [7] to dynamically refresh the graph based on updated node features and layout modality. As for edge features, besides simply concatenating two node features, relationship proposals [8] is also used for better capturing the relative spatial relationship. To improve the computation efficiency and balance between positive and negative training pairs, we also propose a graph sampling method based on layout modality. A sparse graph can benefit forward and backward computations compared to fully-connected graphs. Finally, two linear probes are trained to conduct node classification and edge classification respectively.

Our method does not require the use of a tokenizer or language model to extract text features as part of node embedding, making it language-independent and efficient in terms of parameters. In contrast to Transformer Encoder

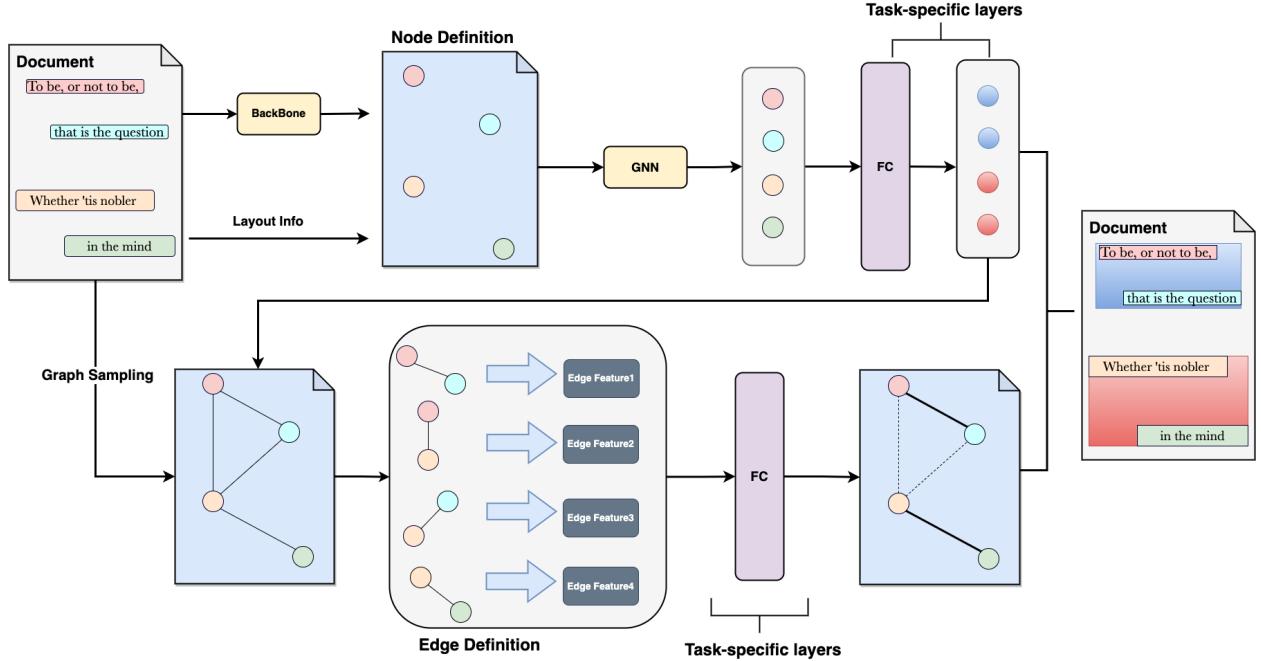


Figure 1: **The overall Paragraph2Graph architecture.** The whole pipeline consists of five parts: node definition, edge definition, GNN, graph sampling and task-specific layers; dotted lines represent invalid edge connections

series or previous GNN works [9][10][11][12][13], we have shown that Paragraph2Graph can easily generalize to multilingual documents without any modifications. We also conducted experiments showing that a model trained on Chinese documents performs similarly or even better on an English evaluation dataset than a model trained on English documents. This demonstrates the language-independence of our approach and indicates that the diversity of document layouts is the primary factor affecting performance. Additionally, our GNN models exhibit better generalization than object detection frameworks such as Faster-RCNN and Mask-RCNN.

Our contributions can be summarized as follows:

- We propose a language-independent GNN framework which we call the Paragraph2Graph. The framework consists of node definition, edge definition, graph sampling, GNN and task-specific layers. Each part of it can be easily edited.
- We offer an empirical selection for each part of the Paragraph2Graph that achieves competitive results on several document layout analysis tasks.
- We conduct extensive experiments and give our ablation analysis to justify the effective design.
- The language-independent design allows us to make use of all public datasets to train a model regardless of language.

2 Related Work

Fig.2 shows three common practices for document layout analysis.

2.1 Layout Tasks use Transformer Encoder

A very-rich document has different modalities available across text info, text position and image. To mimic how humans read, LayoutLM [1], LayoutLMv2 [2], BROS [4] first integrate text features of each token with layout modality and corresponding image features. Afterward, LayoutLMv3 [3] extends the visual backbone to visual transformers. These frameworks define the document layout analysis as a relation extraction and follow the same design by constructing a fully-connected graph to calculate the relation score between all nodes. Then, each relationship in the graph is inferred based on evaluating whether the score is over a threshold or not. All tokens inferred to be relational are grouped as one

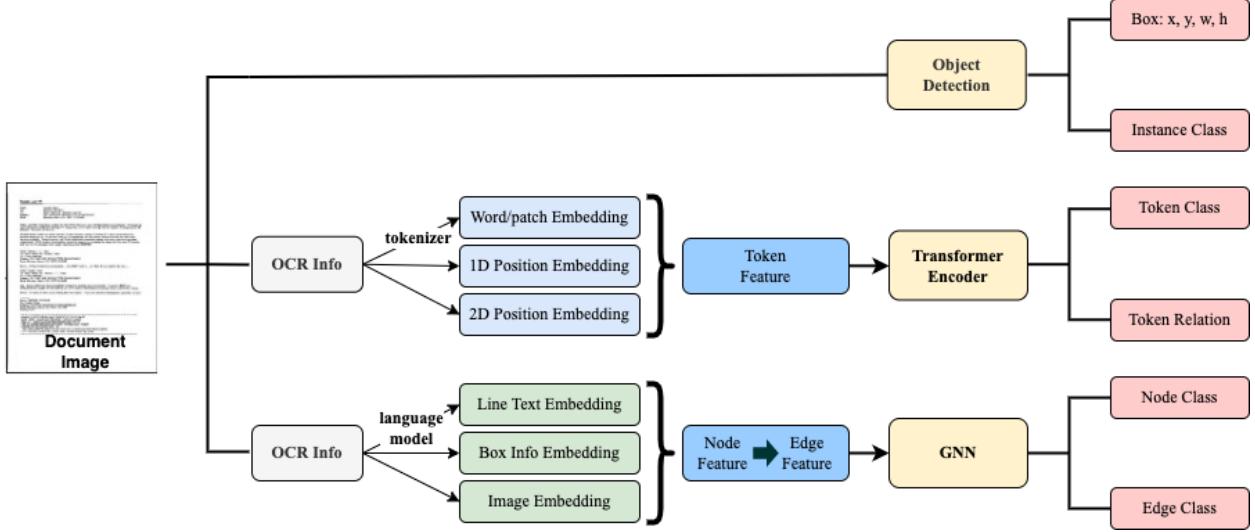


Figure 2: Three common practices for document layout analysis.

region. However, these methods are highly language-dependent because of their language-specific tokenizers. Due to the lack of labeled multilingual document datasets without privacy consideration, it is a challenging alternative plan to train a cross-language tokenizer. Besides, with the limitation of input sequence length, Transformer Encoder-series methods cannot deal with very-long documents, such as dense tables in financial reports. The high computational cost introduced by self-attention and fully-connection graphs reduce the availability to the industry as well.

2.2 Layout Tasks use Object Detection

Document layout analysis is about detecting the layout location of unstructured digital documents by returning bounding boxes and categories such as figures, tables, headers, footers, paragraphs, etc. Such a task is initially defined as an object detection problem on which many algorithms [5][6] [14] related to object detection or segmentation have been successfully applied. However, for all the object detection or segmentation models, the predicted bounding boxes may overlap with each other due to the vague boundary between instances as shown in Fig.3. The slight offset of the prediction boxes has little effect on the training loss, which in turn contributes limited to the model optimization to reach a high IoU. It's hard to assign a label to a text box that is either located at the edge of a predicted region or is shared by multiple predicted regions, which makes the $AP^{IoU \geq 0.9}$ less satisfying. It has to be mentioned that recent works [3][15][16] replacing CNN backbone with vision transformer to achieve state-of-art results on public datasets, but the uncompetitiveness of computation cost can't be ignored.

2.3 Layout Task use GNN

Graph Neural Networks (GNNs) have their special advantages in modeling spatial layout patterns of documents. Each text box of a document can be regarded as a spatially-separated node in a graph; text boxes grouped in the same layout region can be seen as being connected by edges. Since the document layout analysis can be implemented as node and edge classification from a sampled graph, there exist no vague text boxes hard to be assigned to a certain group. We conclude a general pipeline covering all existing GNN-based layout analysis algorithms——node definition, graph sampling, edge definition, GNNs and task-oriented processing. Existing works only explore part of the pipeline.

For node definition, [9] uses char embedding with BiLSTM to integrate text into node; [17] adds regional image feature from FPN output; [10] encodes box coordinate $xywh$ into node embedding. Afterward, Post-OCR [18] adds $\cos\alpha$, $\sin\alpha$, $xcos\alpha$, $xsin\alpha$, $ycos\alpha$, $ysin\alpha$ and the width of the first word, where α is the angle of text box; ROPE[11] explores the importance of the reading orders of given word-level node representations in a graph; Doc2Graph [13] uses a pre-trained U-Net to get text image feature; Doc-GCN [12] proposes a large collection for node definition, which includes text embedding from BERT model, image feature from pre-trained Faster-RCNN model, the number of tokens, the ratio of token number and box area, and syntactic feature.

For edge definition, [9] uses horizontal and vertical distance and the ratio of height between the two text boxes; ROPE[11] constructs with spatial embeddings from horizontal and vertical normalized relative distances between

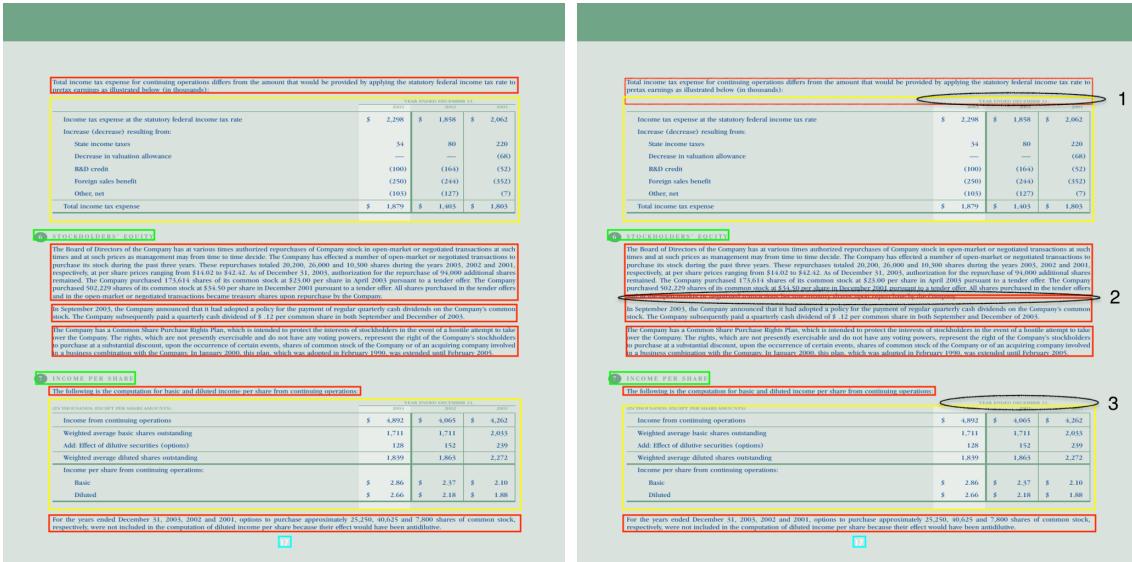


Figure 3: (a) ground truth layout region (b) tricky cases that an object detection method fails to handle with: **region 1**: shows one text box can locate across two layout regions; **region 2**: the text box exactly located at the box boundary; **region 3**: The text box is not located in any regions. green, red, yellow, blue rectangle means detected regions for the Title, Text, Table, and Page Foot.

centers, top left corners, and bottom right corners and relative height and width aspect ratios; Doc2Graph[13] uses the output of the last GNN layer, softmax of the output logits and polar coordinates for node embedding.

For the GNN module, based on the vanilla GNNs, many works have studied sophisticated designs to improve GNNs performances. Graph Convolutional Networks [19] is a type of Graph Neural Network which applies convolution over graph structures. This design is widely used in [12][20]. GAT[21] leverages the self-attention mechanism into GNNs to decouple node update coefficients from the structure of the graph. It has been used in [17][10][18].

For graph sampling, given that a document usually has a large number of text boxes that can be regarded as nodes, it is essential to construct a graph with both high connectivity and sparsity compared to a fully-connected graph to allow necessary gradient propagation. [22] first proposes β -skeleton graph; GraphSage [23] uniformly samples a set of nodes from the neighborhoods and only aggregates feature information from sampled neighbors. [18][11][20] all follow β -skeleton to build their graph, but the miss to cover tabular structures where text box density is relatively high. K-Nearest Neighbor is another good substitution, [10] set $K = 10$ and [24] set $K = 3$, but it is still too tricky to tune satisfying parameters for different business scenarios.

2.4 Other Tasks use GNN: table recognition, text line grouping

All aforementioned algorithms can be generalized to table recognition tasks by simply modifying the task-oriented layers to represent whether two adjacent cells are in the same row or col. For the table recognition task, [25] uses KNN to construct a graph and represent text features by encoding character embedding with GRU; [26] constructs a fully-connection graph and set weighted loss to balance between positive and negative samples.[27] [28] [29] [30] share the identical GNN structure. Text line grouping task is more easily adaptable to GNN with minor changes. [24] predicts edge classified probability to judge if the pivot and its neighbors are in the same line; [31] introduces the residual connection mechanism for GNNs. In general, a powerful GNN model can be used in many downstream document analysis tasks.

3 Method

We follow our conclusion to establish a unified pipeline covering all main steps to build a GNN-based model for layout analysis: they include node definition, graph sampling, edge definition, GNNs, and task-oriented processing.

Node definition Given a document image $D \in \mathbb{R}^{H \times W \times 3}$ with N text boxes generated by any commercial or open source Optical Character Recognition (OCR) engine. we denote all text boxes as: $\text{position_info} = \{x_{min}^n, y_{min}^n, x_{max}^n, y_{max}^n \mid n \in [0, N - 1]\}$. The input image D is first resized into $D' \in \mathbb{R}^{400 \times 400 \times 3}$. D' is sent to a pre-trained ResNet visual backbone to get a series of output features from different scales. These features are integrated with D' into $F \in \mathbb{R}^{400 \times 400 \times d}$ with an FPN structure. For N text boxes, we pick out their corresponding image features with ROIAlign and embed them as $I^{N \times k}$, aka image embedding. d, k is the intermediate dimensions. Given the normalized bounding box, the layout information is represented as $\text{layout}_n = (x_{min}^n, y_{min}^n, x_{max}^n, y_{max}^n, x_{ctr}^n, y_{ctr}^n, w_n, h_n)$. The layout information composes of bounding box coordinates, center point coordinates, bounding box width and height to construct a token-level 2D positional embedding, denoted as layout embedding. We then fuse layout embedding and image embedding:

$$\text{node_embedding} := \text{MLP}(\text{Concat}(\text{image_embedding}, \text{layout_embedding}))$$

GNN module After gathering all the node features, they are passed as input to the interaction model. We have tested two graph neural networks to use as the interaction part which are the modified versions of [7] and [32] respectively. These modified networks are referred to as DGCNN* and GravNet* hereafter. We update the node features by aggregating weighted neighbor nodes with DGCNN/GravNet.

$$\text{node_embedding} = \max(\text{node_embedding}, \text{DGCNN or GravNet}(\text{node_embedding}, \text{position_info}))$$

Graph sampling Text boxes classified into the same layout category can be regarded as having an edge between them. We refer to this task as node grouping: to infer whether there exists an edge between a node pair. To construct node pairs, we therefore connect all potential edges between the nodes with a location-based node search algorithm. Instead of constructing a fully-connected graph, our method can both save computation costs and improve training efficiency. Based on the common structure of a document, each text node can have potential edge connections both vertically and horizontally. For each text box, we pick up its top 1-2 location-nearest text boxes in four directions (top, bottom, left and right). Complex cases need additional processing as shown in Fig.4-b. Our method can effectively sample a sparse graph without missing necessary node pairs. Comparison among KNN, β -skeleton, and our sampling methods can be found in Appendix Fig.5.

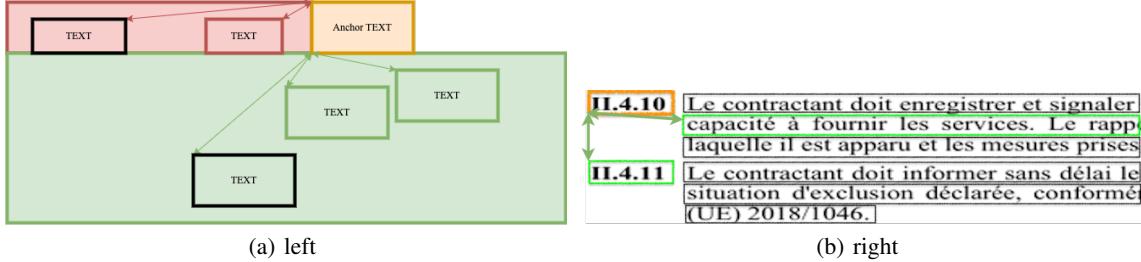


Figure 4: **Left:** illustration of our graph sampling strategy: for each node (shown in orange), we sample one edge horizontally (shown as red region) and two edges vertically (shown as green region); **Right:** We vertically sample top-2 nearest edges instead top-1 to ensure the connectivity with this common right-alignment paragraph structures

Edge definition We concatenate the node feature of each valid node pair as F_{pair} . Inspired by ROPE [11], we encode natural reading orders of words as F_{rope} to help capture the better sequential presentation between nodes. A new reading order code is first assigned to neighbors with respect to each text box. Then, a sinusoidal encoding matrix is applied to encode the reading order index.

We also consider that the relationship between nodes is an important feature that has been ignored by [8]. Following the relationship proposal, suppose two nodes have a potential relationship, we denote one node as S , a subject, the other as O , an object, and the relationship as R . $\Delta(S, O) = (t_x^{SO}, t_y^{SO}, t_w^{SO}, t_h^{SO}, t_x^{OS}, t_y^{OS})$, where

$$t_x^{SO} = (x^S - x^O)/w^S, \quad t_y^{SO} = (y^S - y^O)/h^S \quad (1)$$

$$t_w^{SO} = \log(w^S/w^O), \quad t_h^{SO} = \log(h^S/h^O) \quad (2)$$

$$t_x^{OS} = (x^O - x^S)/w^O, \quad t_y^{OS} = (y^O - y^S)/h^O \quad (3)$$

x^S, y^S, w^S, h^S represent the center coordinates, width, and height of a subject box, similarly denotations apply for x^O, y^O, w^O, h^O . The coordinates of R is the minimum bounding rectangle of S, O , which means

$$(x_{min}^R, y_{min}^R, x_{max}^R, y_{max}^R) = \min(x_{min}^O, x_{min}^S), \min(y_{min}^O, y_{min}^S), \max(x_{max}^O, x_{max}^S), \max(y_{max}^O, y_{max}^S)$$

The relationship feature F_{rel} is defined as an concatenation of $\Delta(S, O)$, $\Delta(S, R)$ and $\Delta(O, R)$. Finally, the edge feature is formally represented as:

$$F_{edge} = \text{Concat}(F_{pair}, F_{rope}, F_{rel})$$

Task-oriented processing For node classification, we apply a fully-connected layer to fuse features and a linear layer $W^{h \times c}$ to classify each node. h is the hidden dimension and c is the number of categories. For node grouping, we follow the same idea as node classification: a fully-connected layer to fuse features and a linear layer $W^{h \times 2}$ to infer whether there exists an edge between a node pair or not. All connected nodes are regarded as layout instance. The minimum bounding boxes of connected nodes are the final layout bounding boxes. The category mode of connected nodes is the category of the layout instance.

4 Experiments

Previous works propose various definitions on edge and nodes, but they either get non-competitive results or introduce expensive computation costs. We therefore study the combinations of these designs and compare them with object detection and the Transformer Encoder model on several document layout analysis tasks. Our experiments demonstrate the effectiveness and competitiveness of our method.

We train our model with 1 GeForce 3090 GPUs from scratch. We use an Adam optimizer with 0.937 momentum and 0.005 weight decay. The learning rate is set to 0.0001.

4.1 Results on Public Datasets

FUNSD The FUNSD [33] provides 199 annotated forms with 9, 707 entities and 31, 485 word-level annotations for four entity types: header, question, answer, and other. It includes noisy scanned documents in English from various fields, such as research, marketing, and advertising. This dataset is commonly used in GNN-related papers, and we used it for our experiments for easy comparison. FUNSD contains two level labels: word and entity. For word-level labels, we predict the category of each word and determine whether two words belong to the same entity. For entity-level labels, we adds two classification heads: one for entity labeling and the other for entity linking, which predicts whether two entities are matched.

We report our best hyperparameter configuration as shown in Tab.?? ours-Large in the ablation experiment of section 4.3. We train the models with a batch size of 2 for 60 epochs and a warm-up period of 10 epochs. The training and validation set are split as provided, with 149 for training and 50 for evaluation.

To evaluate the performance of our method, we use multi-class F1-scores for node classification and binary edge classification F1-scores for grouping or linking, along with corresponding precision and recall values. Our method significantly outperformed previous works. Despite not using a language model, our model had a significantly smaller number of parameters as shown in Tab.1.

In the entity task, our model shown in Tab.2, achieved a F1 score of 0.80575 for entity-labeling and 0.77031 for entity-linking, using 32.98 million parameters. We achieved state-of-the-art results in entity linking, outperforming other GNN models and the Transformer Encoder series. On the entity labeling task, our F1 score was lower than some Transformer Encoder models such as LayoutLMv2, LayoutLMv3, and BROS. This may be due to those methods having more parameters and being pre-trained on several text-image alignment tasks, giving them strong semantic and visual understanding abilities. Despite being trained from scratch with only 149 samples, our model still outperformed BERT, Roberta, and LayoutLM, and achieves significant improvements over most previous GNN works. However, doc2graph performed 1.6% better than our model on the entity labeling task, but it still suffers from the problems associated with language-based GNNs due to its use of a language model. Compared to other state-of-the-art models, our GNN model performed competitively.

PublayNet PublayNet[37] contains research paper images annotated with bounding boxes and polygonal segmentation across five document layout categories: Text, Title, List, Figure, and Table. The official splits contain 335, 703 training images, 11, 245 validation images, and 11, 405 test images. We train our model on the training split and evaluate our model on the validation split following standard practice. We train our models with the batch size of 4 for 5 epochs

	F1	word-labeling	word-grouping	Params
GNN with Language model	FUNSD[33]	-	0.41	340M
GNN with Language model	Named[10]	-	0.65	201M
GNN with Language model	ROPE[11]	0.5722	0.8933	-
GNN	ours-Large	0.68933	0.91483	32.98M

Table 1: Performance for word level on FUNSD.

	F1	entity-labeling	entity-linking	Params
GNN with Language model	FUNSD[33]	0.57	0.04	340M
GNN with Language model	Named[10]	0.64	0.39	201M
GNN	FUDGE[34]	0.6652	0.5662	17M
GNN	Word-FUDGE[34]	0.7221	0.6258	17M
GNN with Language model	Doc2Graph[13]	0.8225	0.5336	6.2M+
Transformer Encoder	BERT-base[35]	0.6026	0.2765	110M
Transformer Encoder	BERT-L[35]	0.6563	0.2911	340M
Transformer Encoder	RoBERTa-base[36]	0.6648		125M
Transformer Encoder	RoBERTa-L[36]	0.7072		355M
Transformer Encoder	LayoutLM[1]	0.7927	0.4586	113M
Transformer Encoder	LayoutLM-L[1]	0.7789	0.4283	343M
Transformer Encoder	LayoutLMv2[2]	0.8276	0.4291	200M
Transformer Encoder	LayoutLMv2-L[2]	0.8420	0.7057	426M
Transformer Encoder	BROS[4]	0.8305	0.7146	138M
Transformer Encoder	BROS-L[4]	0.8452	0.7701	340M
Transformer Encoder	LayoutLMv3[3]	0.9029		133M
Transformer Encoder	LayoutLMv3-L[3]	0.9208		368M
GNN	ours-Large	0.80575	0.77031	32.98M

Table 2: Performance for entity level on FUNSD. Doc2Graph params counts a spaCy model and a pre-trained U-Net on FUNSD besides its own weights.

and warm-up 1 epoch. Because of training resource limitations, we only report our suboptimal configuration shown in Tab.?? ours-Small.

We measure the performance using the mean average precision (mAP) @ intersection over union (IOU) [0.50:0.95] of bounding boxes and report results in Tab.3 with only two categories : Text and Title. Tab.11 shows all categories. Our proposed GNN-based model outperforms several state-of-the-art object detection models. Specifically, our model achieves a mAP of 0.954 and 0.913 for Text and Title detection, with a model size of 77M. Our model shows better performance than Faster-RCNN, Cascade-RCNN and Mask-RCNN regardless of whether they have been pretrained , which sizes ranging from 168M to 538M. Compared to Faster-RCNN-Q and Post-OCR models with mAPs of 0.914 and 0.892 respectively, our GNN-based model achieves higher accuracy in Text category. Fig.6 shows some cases of our algorithm on this dataset.

	mAP	Text	Title	Size
OD	Faster-RCNN[37]	0.910	0.826	-
OD	Mask-RCNN[37]	0.916	0.84	168M
OD-pretrained	Faster-RCNN[UDoc][16]	0.939	0.885	-
OD-pretrained	Mask-RCNN[DiT-base][15]	0.934	0.871	432M
OD-pretrained	Cascade-RCNN[DiT-base][15]	0.944	0.889	538M
OD-pretrained	Cascade-RCNN[layoutlm-v3][3]	0.945	0.906	538M
OD	Faster-RCNN-Q[18]	0.914	-	-
GNN	Post-OCR[18]	0.892	-	-
GNN	ours-Large	0.954	0.913	77M

Table 3: Performance on Publaynet on paragraph categories.OD-pretrained means object model use pretrained CNN.

Doclaynet Doclaynet[38] is a recently released document layout dataset annotated in COCO format. It contains 80863 manually annotated pages from diverse data sources to represent a wide variability in layouts with 11 distinct classes: Caption, Footnote, Formula, List-item, Page-footer, Page-header, Picture, Section-header, Table, Text, and Title. Compared with publaynet, this dataset covers more complex and diverse document types, including Financial Reports, Manuals, Scientific Articles, Laws Regulations, Patents and Government Tenders.

We use the same training parameters as in PubLayNet and evaluate the quality of their predictions using mean average precision (mAP) with 10 overlaps that range from 0.5 to 0.95 in steps of 0.05 (mAP@0.5-0.95). These scores are computed by leveraging the evaluation code provided by the COCO API. Similarly we only compare categories belonging to the Paragraph type without Table and Picture. The results shown in Tab.?? draw a similar conclusion as we do in Publaynet: ours achieves better results with only 1/7 parameters with respect to object detection models in total with a mAP of 0.771. Specifically, our model performs exceptionally well in Page-Header, Caption, Section-Header, Title, and Text detection with mAPs of 0.796, 0.809, 0.824, 0.643, and 0.827, respectively. Comparatively, YOLO-v5x6 achieves best result in Footnote, Title and Text. Tab.12 shows all categories.

mAP	Mask-RCNN-res50[38]	Mask-RCNN-resnext101[38]	Faster-RCNN-resnext101[38]	YOLO-v5x6[38]	ours-Small
Page-Header	0.719	0.7	0.720	0.679	0.796
Caption	0.684	0.715	0.701	0.777	0.809
Formula	0.601	0.634	0.635	0.662	0.726
Page-Footer	0.616	0.593	0.589	0.611	0.920
Section-Header	0.676	0.693	0.684	0.746	0.824
Footnote	0.709	0.718	0.737	0.772	0.625
Title	0.767	0.804	0.799	0.827	0.643
Text	0.846	0.858	0.854	0.881	0.827
Total-paragraph	0.702	0.7143	0.7148	0.744	0.771
Params	-	-	60M	140.7M	19.95M

Table 4: performance on Doclaynet on paragraph categories.

4.2 Discussion on Generalization of Language

To demonstrate the language-independence of our model, we first train our model on datasets in English and evaluate them on a dataset in Chinese. Among them, Publaynet is a pure English dataset, Doclaynet is mostly in English, and DGDoc is a pure Chinese dataset containing 12,000 images from real business scenarios. We use the same F1 indicator as the experiment on FUNSD. As shown in Tab.5, our models trained on Doclaynet data and models trained on Chinese-data behave similarly on Publaynet, and even the latter one gets a higher F1 on two tasks. The model trained on Publaynet data and the model trained on Chinese-data perform similarly on Doclaynet. This experiment proves that our model has the language-independent ability, which allows us to focus more on training data with diversity and complex layout structures instead of languages. The most important advantage of this conclusion is that for non-English application scenarios, we do not need to collect and annotate a large number of documents, which is very time-consuming and expensive. Instead, we can directly collect a variety of public datasets regardless of languages to train the model.

F1	val-doclaynet		val-publaynet		val-chineseData	
	node	edge	node	edge	node	edge
train-doclaynet	0.94670	0.97267	0.95774	0.97171	0.85892	0.94851
train-publaynet	0.72742	0.86069	0.98729	0.99302	0.76158	0.8839
train-chineseData	0.88256	0.92285	0.96289	0.96847	0.97295	0.98872

Table 5: Comparison results for discussion on generalization of language.

4.3 Discussion on Generalization of Data Complexity

In order to compare the generalization of the model we use two datasets: Doclaynet is a more diverse dataset than Publaynet in layout. As shown in Tab.6, if we trained on Publaynet and predicted on Doclaynet, both models Mask-RCNN or our GNN-based model drop badly in mAP. But if we use the model trained on Doclaynet to predict on Publaynet, our model only slightly decreases, while Mask-RCNN drops significantly. It shows that as long as our model has been trained on complex and diverse layouts, it can also migrate to simple layouts well.

mAP		val-publaynet		val-doclaynet	
		maskrcnn-res50	ours-Small	maskrcnn-res50	ours-Small
train-publaynet	Section-Header	0.87	0.913	0.32	0.484
	Text	0.96	0.954	0.42	0.338
train-doclaynet	Section-Header	0.53	0.811	0.68	0.796
	Text	0.77	0.769	0.84	0.827

Table 6: Comparison results for discussion on generalization of data complexity.

4.4 Ablation Experiments

4.4.1 Node-definition

When considering the node definition, we tested whether the box information is 4 values or 8, the backbone tried res-18 and res-50, GNN tried DGCNN and GravNet, whether the GNN input and output are residual connect. The impact of each factor was compared on four tasks for more comprehensive assessment in Tab.7. The addition of residual connections (using res-18 or res-50) generally leads to better performance than non-residual networks. The use of more big backbone (res-50 vs res-18) can improve the results for all tasks. When comparing different GNN models, DGCNN tend to perform better than Gravnet. However the importance of box information is vague.

box-infor	backbone	GNN	GNN-res	WL	WG	EL	EG
4	res-18	D	-	0.65189	0.86821	0.76458	0.67797
8	res-18	D	-	0.63914↓	0.86248↓	0.77573	0.69695
8	res-18	D	+	0.65579	0.86399↓	0.77144	0.70509
8	res-50	D	-	0.66774	0.87487	0.77873	0.71613
8	res-18	G	-	0.58861↓	0.84461↓	0.71397↓	0.60032↓

Table 7: Ablation experiments of node definition on FUNSD: WL, WG, EL, EG means word-labeling, word-grouping, entity-labeling, entity-linking; D and G means DGCNN and Gravnet; box-info 4 means x_{min}, y_{min}, w, h , 8 means $x_{min}, y_{min}, x_{max}, y_{max}, x_{center}, y_{center}, w, h$.

4.4.2 Edge-definition

As for edge definition , we tested four factors: whether to use relationshipROPE polar and node class result as shown in Tab.8. Four tasks is also WL (word-labeling), WG (word-grouping), EL (entity-labeling), and EG (entity-linking). The results indicate that adding relation and node-class edges improves the performance of all metrics except EG. Adding the ROPE edge improves WL and WG but not EL and EG. Adding the polar edge does not significantly affect the results.

relation[8]	ROPE[11]	polar[13]	node-class[13]	WL	WG	EL	EG
			+	0.63478	0.86348	0.77873	0.70071
		+	+	0.63914	0.86248↓	0.77573↓	0.69695↓
+			+	0.63972	0.90039	0.77916	0.74759
+	+	+	+	0.63374↓	0.89869↓	0.77358↓	0.74925

Table 8: Ablation experiments of edge definition on FUNSD: WL, WG, EL, EG means word-labeling, word-grouping, entity-labeling, entity-linking.

4.4.3 Input Process Loss

Tab.9 shows the results of various ablation experiments conducted on the FUNSD dataset, using different input processes and loss functions. The first row represents the baseline model, which uses a 400*400 image size and cross-entropy (CE) loss. The effect of image padding is investigated by adding a padding of unknown size to the input images. The results indicate that this modification does not significantly affect the word-labeling metric but has a negative impact on word-grouping, entity-labeling, and entity-linking. Increasing the image size to 800*608 pixels leads to some

improvement in the word-grouping and entity-labeling metrics, but the word-labeling and entity-linking scores remain relatively low. Adding a contrastive loss term in addition to the cross-entropy loss leads to some improvement in all four tasks. Overall, Padding may not be helpful, larger image sizes may not lead to significant improvements, and the addition of a contrastive loss may be beneficial.

image-size	image-pad	loss	WL	WG	EL	EG
400*400		CE	0.61066	0.86352	0.78216	0.72722
400*400	+	CE	0.63914	0.86248↓	0.77573↓	0.69695↓
800*608	+	CE	0.68795	0.87782	0.81947	0.70487↓
400*400	+	CE+Con	0.64178	0.87105	0.77658↓	0.71704↓

Table 9: Ablation experiments of input process and loss on FUNSD: WL, WG, EL, EG means word-labeling, word-grouping, entity-labeling, entity-linking; CE and Con means CE loss and Contrastive loss.

4.4.4 Final Best Model

Name	backbone	image-size	WL	WG	EL	EG
ours-Small	res-18	400*400	0.66579	0.90148	0.78216	0.75522
ours-Large	res-50	800*608	0.68933	0.91483	0.82504	0.77031

Table 10: Best model.

We study the combinations of a variety of designs for each component in the GNN model and justify their effectiveness on our document layout analysis tasks.

As seen in Tab.10, scale-up image size, relationship proposal, and larger CNN backbone are useful to improve accuracy, while contrast loss, and Gravnet are not. The importance of other factors is ambiguous to tell. DGCNN is a better choice than Gravnet. Based on the detailed ablation experiment, we get the best design combination is ours-Large. Since we enlarge the input image size, more memory is cost. Therefore, on the large datasets DoclayNet and PublayNet, we finally report the suboptimal configuration ours-Small instead.

5 Conclusion Future Work

In this paper, we propose a language-independent GNN framework for document layout analysis tasks. Our proposed model, Paragraph2Graph, uses a pre-trained CNN to encode image features and incorporates 2d OCR text coordinates and image features as node features in a graph. We use a dynamic graph convolutional neural network (DGCNN) to update the graph based on these features and include edge features based on relationships. To improve efficiency, we also propose a graph sampling method based on layout modality. Our method does not require a tokenizer or language model and can easily generalize to multilingual documents without modifications. We show that our method can achieve competitive results on three public datasets with fewer parameters. There are several potential improvements and attempts we leave for future work: (1) We have only experimented with only a few common GNNs, while torch-geometric [39] officially offers nearly 60 related algorithms. Some of them can be a better substitution for DGCNN. (2) The backbone of image features can be pre-trained on document data making it better at capturing document image features. (3) Similar to the layoutLM which has a reasonable pre-training task to improve the merge of different modalities, our model can be pre-trained with image reconstruction tasks as well, such as MAE. (4) Our model doesn't behave well on grouping tables and figures. Future research is needed to expand its generality on these important document layout components.

Acknowledgments

We would like to acknowledge Xinxing Pan, Weihao Li, Binbin Yang Hailong Zhang for their helpful suggestions.

References

- [1] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming 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*, pages 1192–1200, 2020.
- [2] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *ArXiv*, abs/2012.14740, 2020.
- [3] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. *arXiv preprint arXiv:2204.08387*, 2022.
- [4] Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. Bros: A layout-aware pre-trained language model for understanding documents. *ArXiv*, abs/2108.04539, 2021.
- [5] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.
- [6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [7] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019.
- [8] Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed Elgammal. Relationship proposal networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5678–5686, 2017.
- [9] Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. Graph convolution for multimodal information extraction from visually rich documents. *arXiv preprint arXiv:1903.11279*, 2019.
- [10] Manuel Carbonell, Pau Riba, Mauricio Villegas, Alicia Fornés, and Josep Lladós. Named entity recognition and relation extraction with graph neural networks in semi structured documents. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9622–9627, 2021.
- [11] Chen-Yu Lee, Chun-Liang Li, Chu Wang, Renshen Wang, Yasuhisa Fujii, Siyang Qin, Ashok Popat, and Tomas Pfister. Rope: reading order equivariant positional encoding for graph-based document information extraction. *arXiv preprint arXiv:2106.10786*, 2021.
- [12] Siwen Luo, Yihao Ding, Siqu Long, Josiah Poon, and Soyeon Caren Han. Doc-GCN: Heterogeneous graph convolutional networks for document layout analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2906–2916, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.
- [13] Andrea Gemelli, Sanket Biswas, Enrico Civitelli, Josep Lladós, and Simone Marinai. Doc2graph: a task agnostic document understanding framework based on graph neural networks. *arXiv preprint arXiv:2208.11168*, 2022.
- [14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- [15] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Chaoxi Zhang, and Furu Wei. Dit: Self-supervised pre-training for document image transformer. *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.
- [16] Jiuxiang Gu, Jason Kuen, Vlad I. Morariu, Handong Zhao, Nikolaos Barmpalias, R. Jain, Ani Nenkova, and Tong Sun. Unidoc: Unified pretraining framework for document understanding. In *Neural Information Processing Systems*, 2021.
- [17] Johannes Michael, Max Weidemann, Bastian Laasch, and Roger Labahn. Icpr 2020 competition on text block segmentation on a newseye dataset. In *ICPR Workshops*, 2020.
- [18] Renshen Wang, Yasuhisa Fujii, and Ashok Popat. Post-ocr paragraph recognition by graph convolutional networks. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2533–2542, 2021.
- [19] Thomas Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ArXiv*, abs/1609.02907, 2016.
- [20] Shuang Liu, Renshen Wang, Michalis Raptis, and Yasuhisa Fujii. Unified line and paragraph detection by graph convolutional networks. In *DAS*, 2022.

- [21] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
- [22] David G Kirkpatrick and John D Radke. A framework for computational morphology. In *Machine Intelligence and Pattern Recognition*, volume 2, pages 217–248. Elsevier, 1985.
- [23] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.
- [24] Shi-Xue Zhang, Xiaobin Zhu, Jie-Bo Hou, Chang Liu, Chun Yang, Hongfa Wang, and Xu-Cheng Yin. Deep relational reasoning graph network for arbitrary shape text detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9696–9705, 2020.
- [25] Yiren Li, Zheng Huang, Junchi Yan, Yi Zhou, Fan Ye, and Xianhui Liu. Gfte: Graph-based financial table extraction. In *ICPR Workshops*, 2020.
- [26] Wenyuan Xue, Qingyong Li, and Dacheng Tao. Res2tim: Reconstruct syntactic structures from table images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 749–755, 2019.
- [27] Zewen Chi, Heyan Huang, Heng-Da Xu, Houjin Yu, Wanxuan Yin, and Xian-Ling Mao. Complicated table structure recognition. *ArXiv*, abs/1908.04729, 2019.
- [28] Shah Rukh Qasim, Hassan Mahmood, and Faisal Shafait. Rethinking table recognition using graph neural networks. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 142–147, 2019.
- [29] Sachin Raja, Ajoy Mondal, and C. V. Jawahar. Table structure recognition using top-down and bottom-up cues. *ArXiv*, abs/2010.04565, 2020.
- [30] Wenyuan Xue, Baosheng Yu, Wen Wang, Dacheng Tao, and Qingyong Li. Tgrnet: A table graph reconstruction network for table structure recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1275–1284, 2021.
- [31] Pau Riba, Anjan Dutta, Lutz Goldmann, Alicia Fornés, Oriol Ramos Terrades, and Josep Lladós. Table detection in invoice documents by graph neural networks. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 122–127, 2019.
- [32] Shah Rukh Qasim, Jan Kieseler, Yutaro Iiyama, and Maurizio Pierini. Learning representations of irregular particle-detector geometry with distance-weighted graph networks. *The European Physical Journal C*, 79:1–11, 2019.
- [33] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE, 2019.
- [34] Brian L. Davis, B. Morse, Brian L. Price, Chris Tensmeyer, and Curtis Wigington. Visual fudge: Form understanding via dynamic graph editing. In *ICDAR*, 2021.
- [35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.
- [36] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.
- [37] Xu Zhong, Jianbin Tang, and Antonio Jimeno-Yepes. Publaynet: Largest dataset ever for document layout analysis. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022, 2019.
- [38] Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed Samy Nassar, and Peter W. J. Staar. Doclaynet: A large human-annotated dataset for document-layout analysis. *ArXiv*, abs/2206.01062, 2022.
- [39] PyG Team. Pytorch geometric. <https://pytorch-geometric.readthedocs.io/en/latest/>, December 2022.

6 Appendix

6.1 Graph Sampling Comparison

As shown in Fig 5, green lines represent valid pairs connected by algorithms; Red represents missing pairs that should be connected; the sampled results of KNN and β -skeleton are very sensitive to different parameters; For (a):when $\beta=1$, sampled graph miss many pairs, $\beta=0.8$ return enough pairs, but more negative pairs are introduced; Our sampling strategy can reach a balance between sparsity and connectivity.

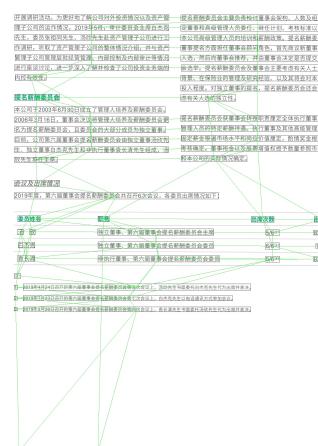


Figure 5: graph sampling results of KNN, β -skeleton and our strategy
13



Figure 6: Result on Publaynet: a) rectangle color red, green, yellow, black means Text, Title, Table, and List; black line means edge between two nodes; b) edge predicted to be connected; c) edge predicted to be unconnected.

mAP	Text	Title	List	Table	Figure	Total	Size
Faster-RCNN[37]	0.910	0.826	0.883	0.954	0.937	0.902	-
Mask-RCNN[37]	0.916	0.84	0.886	0.96	0.949	0.91	168M
Faster-RCNN[UDoc][16]	0.939	0.885	0.937	0.973	0.964	0.939	-
Mask-RCNN[DiT-base][15]	0.934	0.871	0.929	0.973	0.967	0.935	432M
Cascade-RCNN[DiT-base][15]	0.944	0.889	0.948	0.976	0.969	0.945	538M
Cascade-RCNN[layoutlm-v3][3]	0.945	0.906	0.955	0.979	0.970	0.951	538M
Faster-RCNN-Q[18]	0.914						-
Post-OCR[18]	0.892						-
ours-Small	0.954	0.913	0.805	0.932	0.777	0.876	77M

Table 11: performance on Publaynet on all categories.

mAP	Mask-RCNN-res50[38]	Mask-RCNN-resnext101[38]	Faster-RCNN-resnext101[38]	YOLO-v5x6[38]	ours-Small
Page-Header	0.719	0.7	0.720	0.679	0.796
Caption	0.684	0.715	0.701	0.777	0.809
Formula	0.601	0.634	0.635	0.662	0.726
Page-Footer	0.616	0.593	0.589	0.611	0.920
Section-Header	0.676	0.693	0.684	0.746	0.824
Footnote	0.709	0.718	0.737	0.772	0.625
Title	0.767	0.804	0.799	0.827	0.643
Text	0.846	0.858	0.854	0.881	0.827
List-Item	0.812	0.808	0.810	0.862	0.805
Picture	0.717	0.727	0.720	0.771	0.581
Table	0.822	0.829	0.822	0.863	0.559
Total-paragraph	0.702	0.7143	0.7148	0.744	0.771
Total	0.724	0.735	0.734	0.768	0.738
Params	-	-	60M	140.7M	19.95M

Table 12: performance on Doclaynet on all categories.