文献汇报

河北大学
网络空间安全与计算机学院
人工智能
XX
XXXXXXXXXX

VSR: A Unified Framework for Document Layout Analysis combining Vision, Semantics and Relations

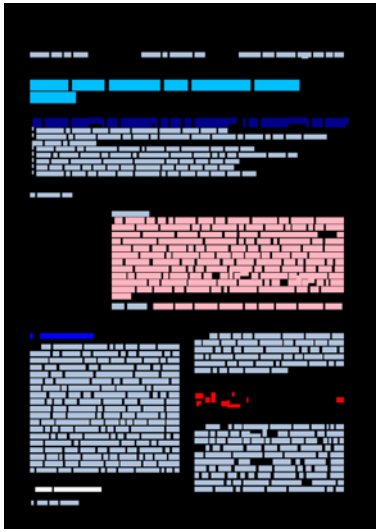
Peng Zhang[1], Can Li[1], Liang Qiao[1], Zhanzhan Cheng[2,1], Shiliang Pu[1], Yi Niu[1], and Fei Wu[2]
[1] Hikvision Research Institute, China
{**zhangpeng23, lican9, qiaoliang6, chengzhanzhan, pushiliang.hri, niuyi**} @hikvision.com
[2] Zhejiang University, China
wufei@cs.zju.edu.cn

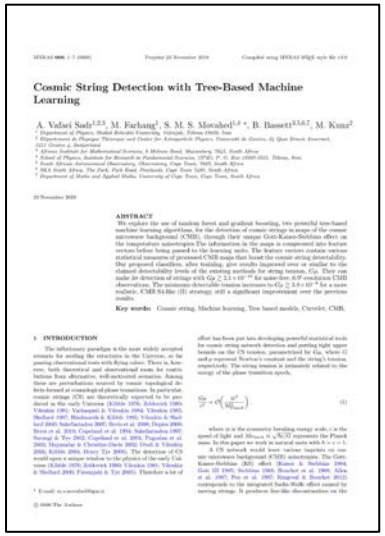
PDF：

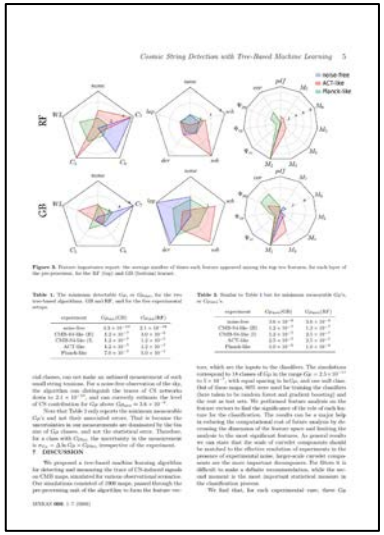
Code：

Question: Layout Analysis

版面分析是指分析一副文本图像的块结构，将图像分成若干具有相似性质的区域（块）以便进行后续的 OCR 识别处理。







Abstract | Author | Caption | Equation | Figure | Footer | List | Paragraph | Reference | Section | Table | Title

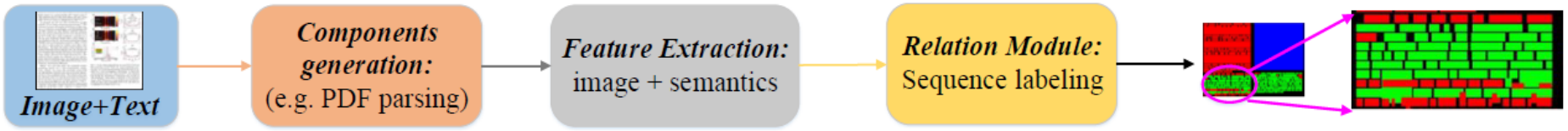
Abstract

Document layout analysis is crucial for understanding document structures. On this task, vision and semantics of documents, and relations between layout components contribute to the understanding process. Though many works have been proposed to exploit the above information, they show unsatisfactory results. NLP-based methods model layout analysis as a sequence labeling task and show insufficient capabilities in layout modeling. CV-based methods model layout analysis as a detection or segmentation task, but bear limitations of inefficient modality fusion and lack of relation modeling between layout components. To address the above limitations, we propose a unified framework VSR for document layout analysis, combining vision, semantics and relations.VSR supports both NLP-based and CV-based methods. Specifically, we first introduce vision through document image and semantics through text embedding maps. Then, modality-specific visual and semantic features are extracted using a two-stream network, which are adaptively fused to make full use of complementary information. Finally, given component candidates, a relation module based on graph neural network is incorported to model relations between components and output final results. On three popular benchmarks, VSR outperforms previous models by large margins. Code will be released soon.

1 Introduction

Many deep learning models have been proposed on this task in both computer vision (CV) and natural language processing (NLP) communities.

对于 Layout Analysis，主要有两派：CV 派和 NLP 派。

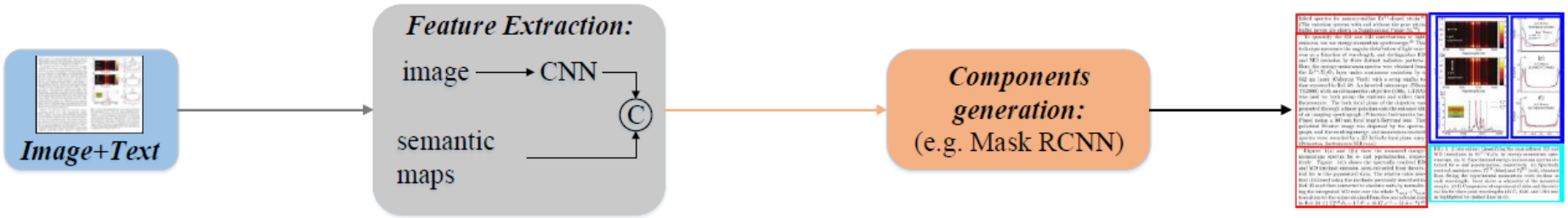


(a) NLP-based framework

NLP-based methods model layout analysis as a sequence labeling task and apply a bottom-up strategy. They first serialize texts into 1D token sequence. Then, using both semantic and visual features (such as coordinates and image embedding) of each token, they determine token labels sequentially through a sequence labeling model. However, NLP-based methods show insufficient capabilities in layout modeling.

NLP 派利用到了语义特征，但是布局建模乱七八糟，不太好使。

1 Introduction



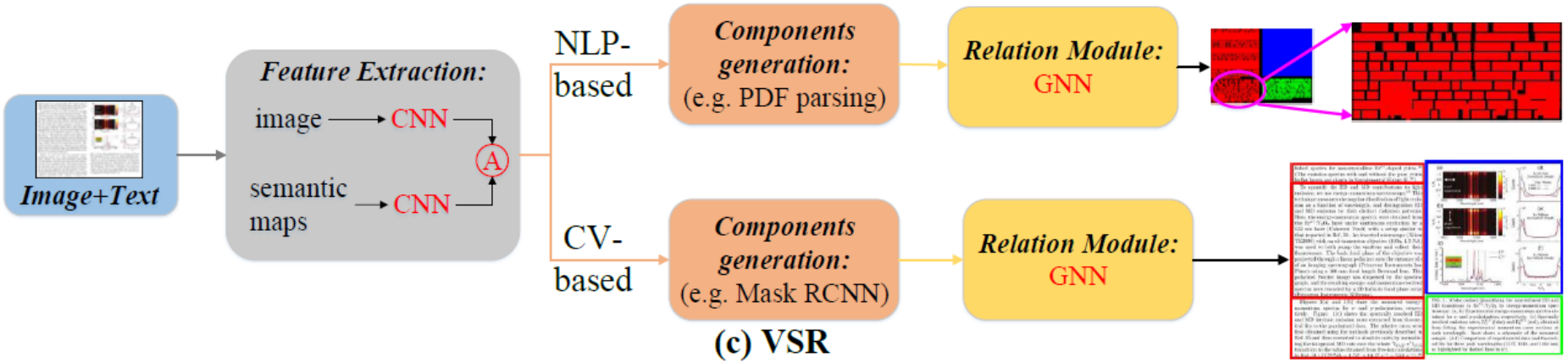
(b) CV-based framework

CV-based methods model layout analysis as object detection or segmentation task, and apply a top-down strategy.
While capturing spatial information better compared to NLP-based methods, CV-based methods still have 3 limitations:

(1) limited semantics
(2) simple and heuristic modality fusion strategy.
(3) lack of relation modeling between components.

CV 派，捕捉空间信息的性能更好，但是不能充分利用语义信息，模态融合太简单，组件间缺乏关系建模，好使但没那么好使。

1 Introduction



(c) VSR

In this paper, we propose a unified framework **VSR** for document layout analysis, combining **Vision**, **Semantics** and **Relation modeling**.This framework can be applied to both NLP-based and CV-based methods.

我们这篇论文提出了一个被命名为 **VSR** 的框架，结合了**视觉**、**语义**和**关系建模**，真是太棒了！

2 Related Works

Document Layout Analysis

- 其他人的单模态布局严格限制于视觉或语义特征，无法利用来自其他模态的补充信息。
- 其他人的多模态布局太简单了，我们提出的高级。

Two-stream networks

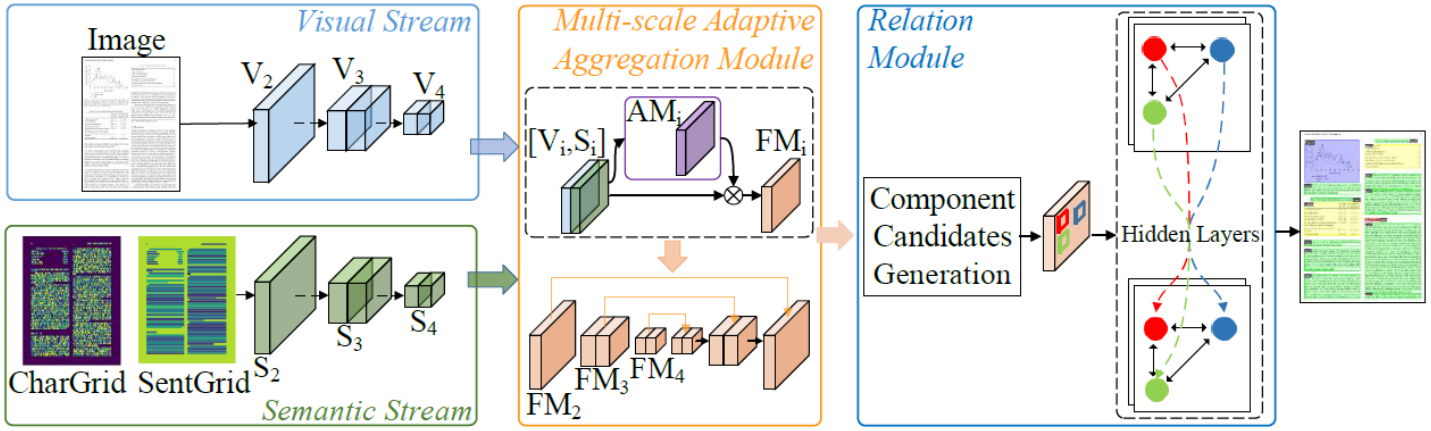
- 双流网络适合用于多模态，用了！

Relation modeling

- NLP 有关系建模，CNN 很难，我们设计了一个 GNN 支持基于 NLP 或 CV 的方法中的关系建模。

3 Methodology

3.1 Architecture Overview



Our proposed framework has three parts: two-stream ConvNets, a multi-scale adaptive aggregation module and a relation module.

我们提出的框架由三个部分：

双流卷积神经网络

- 视觉流（Visual Stream）处理图像，生成检测或分割模型生成的视觉（Visual）特征（如 Mask RCNN）
- 语义流（Semantic Stream）处理文本，生成语义（Semantic）特征（如文本标记）

多尺度自适应聚合模块（Multi-scale Adaptive Aggregation Module）

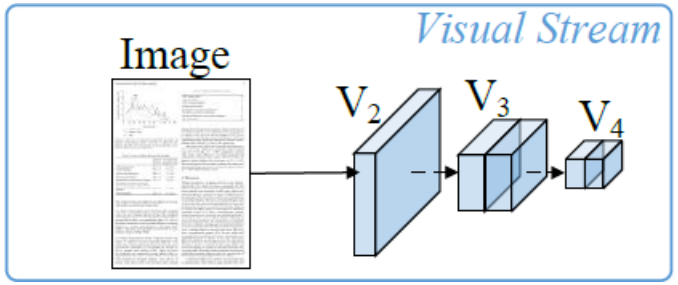
- 聚合视觉特征和语义特征，产生一组候选组件（component candidates）

关系模块

- 接受这些候选组件，生成最终结果（final results）

3 Methodology

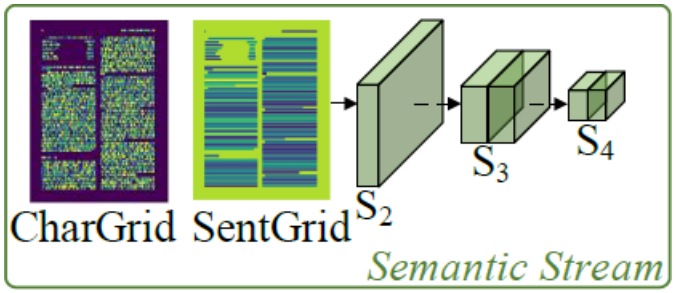
3.2 Two-stream ConvNets



This stream directly takes document images as input and extracts multi-scale deep features using CNN backbones like ResNet.

Visual Stream 直接将文档图像作为输入，并使用 CNN 主干（如 ResNet 提取多尺度深度特征），一阵卷生成视觉特征 $\{V_2, V_3, V_4, V_5\}$。

Semantic Stream 把文本嵌入映射 S_0 作为卷积层的输入一阵卷，最后生成语义特征 $\{S_2, S_3, S_4, S_5\}$。而 S_0 又由 CharGrid 和 SentGrid 组成：$S_0 =$ LayerNorm(CharGrid+SentGrid)，就是字符级和句子级的语义特征。

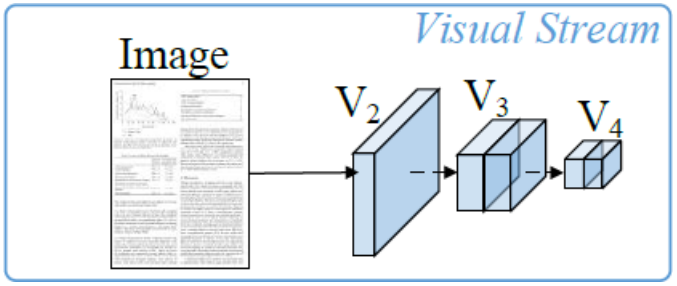


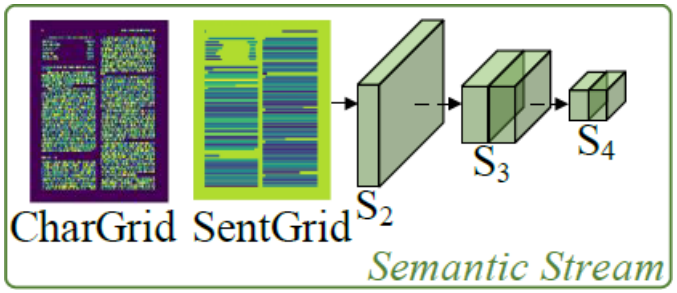
E_c 就是个词嵌入层，将输入层的向量作为输入，将每个词转化成一个更高维度的向量，以便模型可以更好地处理这些信息。

使用训练好的 BERT 模型作为 E_s。

3 Methodology

3.2 Two-stream ConvNets





Train From Scratch

If you want to re-implement the model's performance from scratch, please following these steps:

1.Firstly, prepare the pretrained models:

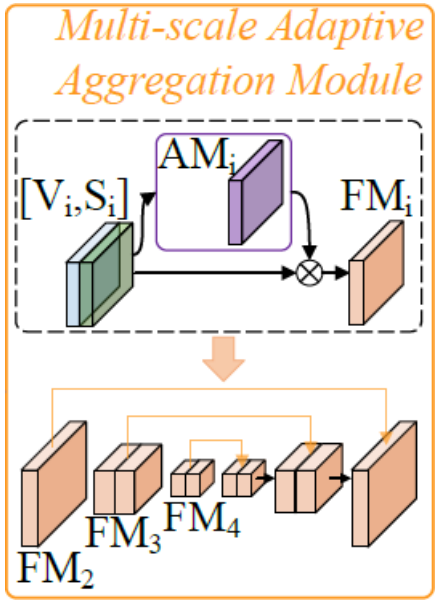
- pretrained mask-rcnn model (Access Code：KZgm) on COCO (*we just copy the params of backbone to initialize backbone_semantic*)
- bert-base-uncased

2.Secondly, modify the paths in model config (`demo/text_layout/VSR/PubLayNet/config/publaynet_x101.py` or `demo/text_layout/VSR/DocBank/config/docbank_x101.py`.), including the pretrained models paths, images paths, work space, etc.

3.Thirdly, direct run `demo/text_layout/VSR/PubLayNet/dist_train.sh` or `demo/text_layout/VSR/DocBank/dist_train.sh`.

3 Methodology

3.3 Multi-scale Adaptive Aggregation



在 i 尺度下，首先将 V_i 和 S_i 连接起来，然后卷一下 $g()$ 再激活 $h()$ 得到 attention map AM_i：$AM_i = h(g([V_i, S_i]))$

然后再一阵操作得到聚合的多模态特征：

$$FM_i = AM_i \odot V_i + (1 - AM_i) \odot S_i$$

得到 $FM = \{FM_2, FM_3, FM_4, FM_5\}$，扔 FPN（特征金字塔网络）中继续卷，提供增强的表示。

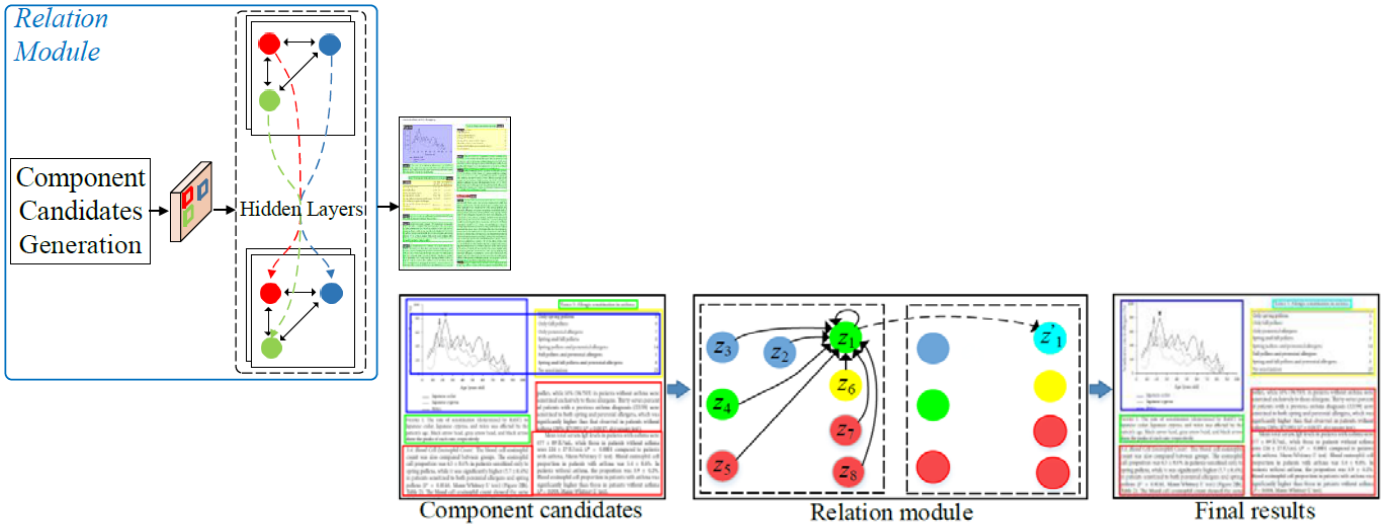
3 Methodology

3.4 Relation Module

其他人到这边就结束了？它们忽视了布局组件之间很强的关系。

布局组件之间存在很强的关系。例如，同列段落的包围框应该对齐；表和表标题经常一起出现；组件之间没有重叠。利用这种关系可以进一步优化。

使用 GNN 建模组件之间的关系，将文档视作一个图 G=(O, E)，O 为节点集（表示之前目标检测模型生成的候选组件），E 是边集（两个候选组件之间的关系）。



节点集的特征 z_j 包含位置坐标 b_j 和深度特征 f_j：$z_j = \mathrm{LayerNorm}(f_j + e_j^{pos}(b_j))$ 为第 j 节点的位置嵌入向量。

节点间的关系用自注意力机制来表示：$\hat{O} = Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d^k}}V)$，$Q,K,V$ 从节点特征集 $Z\{z_1,\dots,z_N\}$ 里整。

一阵操作更新出新的节点特征 Z' 和节点 $\tilde{o_j} = (\widetilde{p_j^c}, \tilde{b_j})$。

$\widetilde{p_j^c} = Softmax(Linear_{cls}(z'_j))$，表示这个节点属于 c 类的概率。

$\tilde{b_j}$ 就是其回归坐标。

3.5 Optimization

我们这个模型可以建模为**序列标记（sequence labeling）**或**对象检测（object detection）**的任务，所以损失函数也不同。

sequence labeling

$$\mathcal{L} = -\frac{1}{T}\sum_{j=1}^{T} \log \tilde{p}_j(y_j)$$

感觉就是交叉熵损失函数?

Layout analysis as object detection

$$\mathcal{L} = \mathcal{L}_{DET} + \lambda \mathcal{L}_{RM}$$

既要看分类的对不对，还要看框的准不准，至于哪个更重要？自行设置超参数 λ。

4 Experiments

4.1 Datasets

Article Regions consists of 822 document samples and 9 region classes are annotated (Title, Authors, Abstract, Body, Figure, Figure Caption, Table, Table Caption and References).

PubLayNet is a large-scale document dataset recently released by IBM. It consists of 360K document samples and 5 region classes are annotated (Text, Title, List, Figure, and Table).

Field	Value
Format	JPG, JSON
License	CDLA-Permissive
Domain	Computer Vision
Number of Records	358,353 images
Data Split	335,703 training images 11,245 validation images 11,405 test images
Size	102 GB

File	Size
DocBank_500K_txt.zip	3,167,771,976B (2.95GB)
DocBank_500K_ori_img.zip [1] [2] [3] [4] [5] [6] [7] [8] [9] [10]	50,907,670,187B (47.4GB)
MSCOCO_Format_Annotation.zip	208,973,824B (199 MB)

DocBank is proposed by Microsoft. It contains 500K document samples with 12 region classes (Abstract, Author, Caption, Equation, Figure, Footer, List, Paragraph, Reference, Section, Table and Title).

4 Experiments

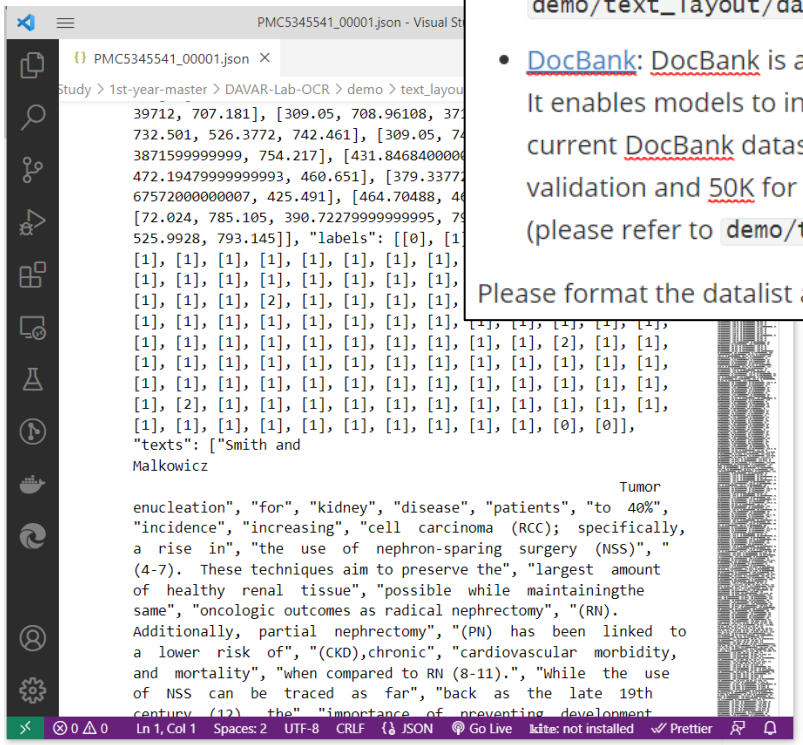
4.1 Datasets

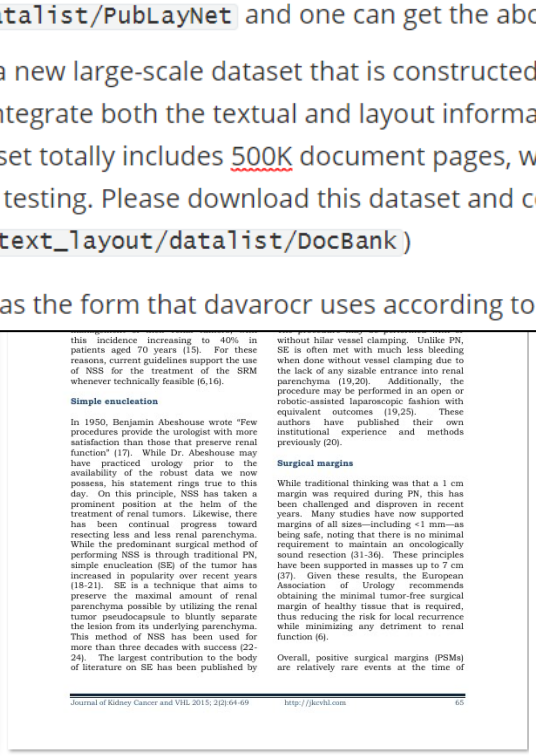
Dataset Preparation

The demos are conducted on two public datasets: PubLayNet and DocBank. Due to the policy, you should download the original data and annoations from the official websites.

- PubLayNet: PubLayNet is a large dataset of document images, of which the layout is annotated with both bounding boxes and polygonal segmentations. To perform multimodal layout analysis task, we also need to extract the annotations in the **character** granularity, in addition to layout component granularity. We provide demo examples in `demo/text_layout/datalist/PubLayNet` and one can get the above annotations through:
- DocBank: DocBank is a new large-scale dataset that is constructed using a weak supervision approach. It enables models to integrate both the textual and layout information for downstream tasks. The current DocBank dataset totally includes 500K document pages, where 400K for training, 50K for validation and 50K for testing. Please download this dataset and convert annotations to *Davar* format (please refer to `demo/text_layout/datalist/DocBank`)

Please format the datalist as the form that davarocr uses according to instructions.




this incidence increasing to 40% in patients aged 70 years (15). For these reasons, current guidelines support the use of NSS for the treatment of the SRM whenever technically feasible (6,16).

Simple enucleation

In 1950, Benjamin Abeshouse wrote "Few procedures provide the urologist with more satisfaction than those that preserve renal function" (17). While Dr. Abeshouse may have practiced urology prior to the availability of the robust data we now possess, his statement rings true to this day. On this principle, NSS has taken a prominent position at the helm of the treatment of renal tumors. Likewise, there has been continual progress toward resecting less and less renal parenchyma. While the predominant surgical method of performing NSS is through traditional PN, simple enucleation (SE) of the tumor has increased in popularity over recent years (18-21). SE is a technique that aims to preserve the maximal amount of renal parenchyma possible by utilizing the renal tumor pseudocapsule to bluntly separate the lesion from its underlying parenchyma. This method of NSS has been used for more than three decades with success (22-24). The largest contribution to the body of literature on SE has been published by

without hilar vessel clamping. Unlike PN, SE is often met with much less bleeding when done without vessel clamping due to the lack of any sizable entrance into renal parenchyma (19,20). Additionally, the procedure may be performed in an open or robotic-assisted laparoscopic fashion with equivalent outcomes (19,25). These authors have published their own institutional experience and methods previously (20).

Surgical margins

While traditional thinking was that a 1 cm margin was required during PN, this has been challenged and disproven in recent years. Many studies have now supported margins of all sizes—including <1 mm—as being safe, noting that there is no minimal requirement to maintain an oncologically sound resection (31-36). These principles have been supported in masses up to 7 cm (37). Given these results, the European Association of Urology recommends obtaining the minimal tumor-free surgical margin of healthy tissue that is required, thus reducing the risk for local recurrence while minimizing any detriment to renal function (6).

Overall, positive surgical margins (PSMs) are relatively rare events at the time of

4 Experiments

4.2 Implementation Details

Document image is directly used as input for visual stream. For semantic stream, we extract embedding maps (*SentGrid* and *CharGrid*) from text as input, where *SentGrid* is generated by pretrained BERT model [8] and *CharGrid* is obtained from a word embedding layer. They all have the same channel dimension size ($C_0^S = 64$). ResNeXt-101 [37] is used as backbone to extract both visual and semantic features (unless otherwise specified), which are later fused by a multi-scale adaptive aggregation and feature pyramid network.

For CV-based multimodal layout analysis methods, fused features are fed into RPN, followed by RCNN, to generate component candidates. In RPN, 7 anchor ratios (0.02, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0) are adopted to handle document elements that vary in sizes and scales. In relation module, dimension of each candidate is set to 1024 and 2 layers of multi-head attention with 16 heads are used to model relations. We set λ in Eq.(11) to be 1 in all our experiments. For NLP-based multimodal layout analysis methods, low-level elements parsed from PDFs (*e.g.*, tokens) serve as component candidates, and relation module predicts their semantic labels.

Our model is implemented under the PyTorch framework. It is trained by the SGD optimizer with batchsize=2, momentum=0.9 and weight-decay=10^{-4}. The initial learning rate is set to 10^{-3}, which is divided by 10 every 10 epochs on Article Regions dataset and 3 epochs on the other two benchmarks. The training of model on Article Regions lasts for 30 epochs while on the other two benchmarks lasts for 6. All the experiments are carried out on Tesla-V100 GPUs. Source code will be released in the near future.

```
# visual branch
backbone=dict(
    type='ResNeXt',
    depth=101,
    groups=64,
    base_width=4,
    num_stages=4,
    out_indices=(0, 1, 2, 3),
    frozen_stages=1,
    norm_cfg=dict(type='BN', requires_grad=True),
    style='pytorch'),

rpn_head=dict(
    type='RPNHead',
    in_channels=256,
    feat_channels=256,
    anchor_generator=dict(
        type='AnchorGenerator',
        scales=[8],
        ratios=[0.02, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0],
        strides=[4, 8, 16, 32, 64]),
    bbox_coder=dict(
        type='DeltaXYWHBBoxCoder',
        target_means=[.0, .0, .0, .0],
        target_stds=[1.0, 1.0, 1.0, 1.0]),
    loss_cls=dict(
        type='CrossEntropyLoss', use_sigmoid=True, loss_weight=
    loss_bbox=dict(type='SmoothL1Loss', beta=1.0 / 9.0, loss_we
```

```
# optimizer
optimizer = dict(type='SGD', lr=0.01, momentum=0.9, weight_decay=0.0001)
optimizer_config = dict(grad_clip=None)
```

4 Experiments

4.3 Results

Trained Model Download

All of the models are re-implemented and well trained based on the opensourced framework mmdetection. So, the results might be slightly different from reported results.

Trained models can be download as follows:

Dataset	Backbone	Pretrained	Test Scale	AP	Links
PubLayNet (Reported)	ResNext101	COCO	(1300, 800)	95.7	-
PubLayNet	ResNext101	COCO	(1300, 800)	95.8	config, pth (Access Code: 9UYK)
DocBank (Reported)	ResNext101	COCO	(600, 800)	95.59	-
DocBank	ResNext101	COCO	(600, 800)	95.25	config, pth (Access Code: IJsy)

Table 1. Performance comparisons on Article Regions dataset

Method	Title	Author	Abstract	Body	Figure	Figure Caption	Table	Table Caption	Reference	mAP
Faster RCNN [31]	-	1.22	-	87.49	-	-	-	-	-	46.38
Faster RCNN *w/ context* [31]	-	10.34	-	93.58	-	-	-	30.8	-	70.3
Faster RCNN *reimplement*	100.0	51.1	94.8	98.9	94.2	91.8	97.3	67.1	90.8	87.3
Faster RCNN *w/ context reimplement* [31]	100.0	60.5	90.8	98.5	**96.2**	91.5	**97.5**	64.2	91.2	87.8
VSR	**100.0**	**94**	**95**	**99.1**	95.3	**94.5**	96.1	**84.6**	**92.3**	**94.5**

Note: missing entries are because those results are not reported in their original papers.

Table 2. Performance comparisons on PubLayNet dataset.

Method	Dataset	Text	Title	List	Table	Figure	AP
Faster RCNN [43]	val	91	82.6	88.3	95.4	93.7	90.2
Mask RCNN [43]		91.6	84	88.6	96	94.9	91
VSR		**96.7**	**93.1**	**94.7**	**97.4**	**96.4**	**95.7**
Faster RCNN [43]	test	91.3	81.2	88.5	94.3	94.5	90
Mask RCNN [43]		91.7	82.8	88.7	94.7	95.5	90.7
DocInsightAI		94.51	88.31	94.84	95.77	97.52	94.19
SCUT		94.3	89.72	94.25	96.62	97.68	94.51
SRK		94.65	89.98	**95.14**	**97.16**	**97.95**	94.98
SiliconMinds		96.2	89.75	94.6	96.98	97.6	95.03
VSR		**96.69**	**92.27**	94.55	97.03	97.90	**95.69**

Table 3. Performance comparisons on DocBank dataset in F1 Score.

Method	Abstract	Author	Caption	Equation	Figure	Footer	List	Paragraph	Reference	Section	Table	Title	Macro Average
$BERT_{base}$	92.94	84.84	86.29	81.52	100.0	78.05	71.33	96.19	93.10	90.81	82.96	94.42	87.70
$RoBERTa_{base}$	92.88	86.18	89.44	82.48	100.0	80.14	73.53	96.46	93.41	93.37	83.89	95.11	88.91
$LayoutLM_{base}$	98.16	85.95	95.97	89.47	100.0	89.57	89.48	97.88	93.38	95.98	86.33	95.79	93.16
$BERT_{large}$	92.86	85.77	86.50	81.77	100.0	78.14	69.60	96.19	92.84	90.65	83.20	94.30	87.65
$RoBERTa_{large}$	94.79	87.24	90.81	83.70	100.0	83.92	74.51	96.65	93.34	94.07	84.94	94.61	89.88
$LayoutLM_{large}$	97.84	87.83	95.56	89.74	**100.0**	91.46	90.04	97.90	93.32	95.96	86.79	95.52	93.50
X101	97.17	82.27	94.35	89.38	88.12	90.29	90.51	96.82	87.98	94.12	83.53	91.58	90.51
$X101+LayoutLM_{base}$	98.15	89.07	**96.69**	94.30	99.90	92.92	93.00	98.43	94.37	96.64	88.18	95.75	94.78
$X101+LayoutLM_{large}$	98.02	89.64	96.66	94.40	99.94	93.52	92.93	98.44	94.30	96.70	88.75	95.31	94.88
VSR	**98.29**	**91.19**	96.32	**95.84**	99.96	**95.11**	**94.66**	**98.66**	**95.05**	**97.11**	**89.24**	**95.63**	**95.59**

Table 4. Performance comparisons on DocBank dataset in mAP.

Models	Abstract	Author	Caption	Equation	Figure	Footer	List	Paragraph	Reference	Section	Table	Title	mAP
Faster RCNN	96.2	88.9	93.9	**78.1**	85.4	**93.4**	86.1	67.8	89.9	76.7	77.2	**95.3**	86.3
VSR	**96.3**	**89.2**	**94.6**	77.3	**97.8**	93.2	**86.2**	**69.0**	**90.3**	**79.2**	**77.5**	94.9	**87.6**

4 Experiments

4.4 Ablation Studies

Table 5. Effects of semantic features at different granularities.

Vision	Semantics		Title	Author	Abstract	Body	Figure	Figure Caption	Table	Table Caption	Reference	mAP
	Char	Sentence										
√			100.0	51.1	94.8	98.9	94.2	91.8	97.3	67.1	90.8	87.3
√	√		100.0	71.4	**96.5**	98.9	95.6	**93.6**	96.9	68.6	89.9	90.2
√		√	100.0	60.2	95.5	**99.0**	**97.8**	93.2	98.9	**73.0**	91.2	89.8
√	√	√	**100.0**	**84.3**	96.1	98.7	95.7	92.5	**99.4**	71.4	**92.4**	**92.3**

Table 6. Effects of two-stream network with adaptive aggregation.

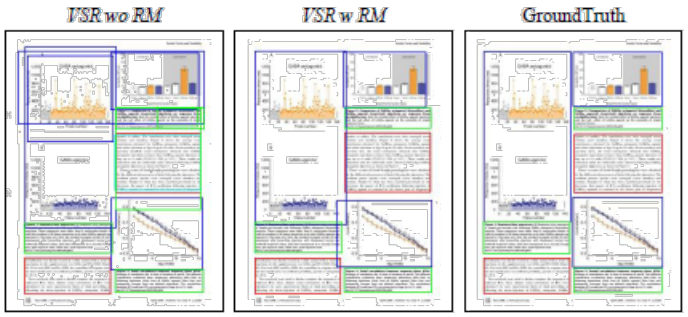
Method		Title	Author	Abstract	Body	Figure	Figure Caption	Table	Table Caption	Reference	mAP	FPS
Single-stream at input level	R101	94.7	58.7	82.7	98.1	97.9	**96.3**	91.8	63.7	91.5	86.2	19.07
	R152	100.0	50.5	85.3	97.9	**98.0**	94.4	93.3	62.6	90.5	85.8	18.15
Single-stream at decision level	R101	99.5	67.6	95.1	98.8	95.0	93.2	96.6	70.7	91.3	89.8	**19.79**
	R152	100.0	80.2	91.0	**99.4**	96.0	92.4	98.3	**73.8**	91.7	91.4	16.43
VSR	R101	**100.0**	**84.3**	**96.1**	98.7	95.7	92.5	**99.4**	71.4	**92.4**	**92.3**	13.94

我们这个模型设计的还是满有效的！

Table 7. Effects of relation module.

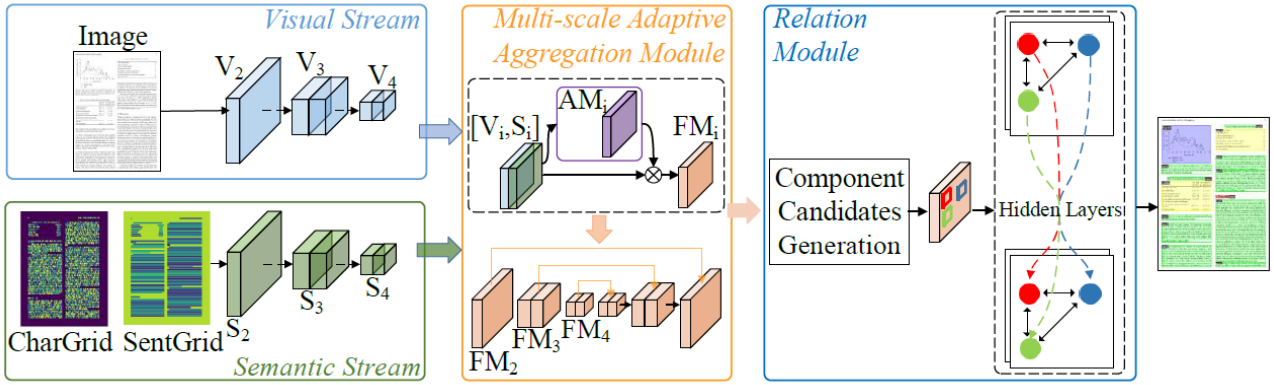
Method		Title	Author	Abstract	Body	Figure	Figure caption	Table	Table caption	Reference	mAP
Faster RCNN	w/o RM	1	51.1	94.8	98.9	**94.2**	91.8	97.3	67.1	90.8	87.3
	w/ RM	**1**	**88.4**	**99.1**	**99.1**	85.4	**92.6**	**98.0**	**79.2**	**91.6**	**92.6**
VSR	w/o RM	1	84.3	**96.1**	98.7	**95.7**	92.5	**99.4**	71.4	**92.4**	92.3
	w/ RM	**1**	**94**	95	**99.1**	95.3	**94.5**	96.1	**84.6**	92.3	**94.5**

Limitations. As mentioned above, in addition to document images, VSR also requires the positions and contents of texts in the document. Therefor, the generalization of VSR may be not good enough compared with its unimodal counterparts, which we'll address in the future.



5 Conclusion

In this paper, we present a unified framework VSR for multimodal layout analysis combining vision, semantics and relations. We first introduce semantics of document at character and sentence granularities. Then, a two-stream convolutional network is used to extract modality-specific visual and semantic features, which are further fused in the adaptive aggregation module. Finally, given component candidates, a relation module is adopted to model relations between them and output final results. On three benchmarks, VSR outperforms its unimodal and multimodal single-stream counterparts significantly. In the future, we will investigate pre-training models with VSR and extend it to other tasks, such as information extraction.



这个模型用双流卷积网络提取了视觉和语义特征，把这些特征在自适应聚合模块中融合，最后使用关系模块进行建模，好使！

拜拜