GeoLayoutLM: Geometric Pre-training for Visual Information Extraction

Chuwei Luo*, Changxu Cheng*, Qi Zheng, Cong Yao
DAMO Academy, Alibaba Group
{luochuwei,ccx0127,zhengqisjtu,yaocong2010}@gmail.com

Abstract

*Visual information extraction (VIE) plays an important role in Document Intelligence. Generally, it is divided into two tasks: semantic entity recognition (SER) and relation extraction (RE). Recently, pre-trained models for documents have achieved substantial progress in VIE, particularly in SER. However, most of the existing models learn the geometric representation in an implicit way, which has been found insufficient for the RE task since geometric information is especially crucial for RE. Moreover, we reveal another factor that limits the performance of RE lies in the objective gap between the pre-training phase and the fine-tuning phase for RE. To tackle these issues, we propose in this paper a multi-modal framework, named GeoLayoutLM, for VIE. GeoLayoutLM explicitly models the geometric relations in pre-training, which we call geometric pre-training. Geometric pre-training is achieved by three specially designed geometry-related pre-training tasks. Additionally, novel relation heads, which are pre-trained by the geometric pre-training tasks and fine-tuned for RE, are elaborately designed to enrich and enhance the feature representation. According to extensive experiments on standard VIE benchmarks, GeoLayoutLM achieves highly competitive scores in the SER task and significantly outperforms the previous state-of-the-arts for RE (*e.g.*, the F1 score of RE on FUNSD is boosted from 80.35% to 89.45%)* [1].

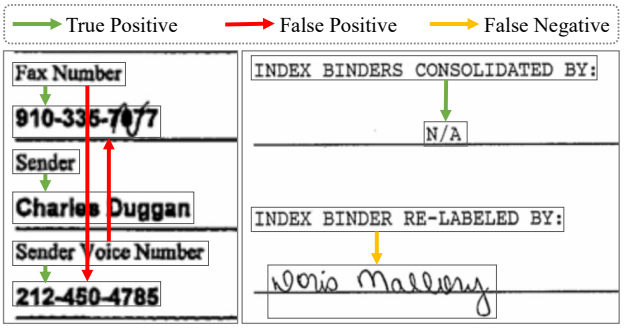


Figure 1. Incorrect relation predictions by the previous state-of-the-art model LayoutLMv3 [15]. (a) LayoutLMv3 tends to link two entities relying more on their semantics than the geometric layout, *i.e.*, the entity "212-450-4785" is linked to "Fax Number" regardless of their relationship in layout. (b) LayoutLMv3 successfully predicts the link in the upper half part but misses the link below, although both links are similar in geometric layout. These two examples clearly show **the importance of geometric information in relation extraction (RE)**.

	Precision	Recall	F1
LayoutLMv3	75.82	85.45	80.35
+ geometric constraint	**79.87**	85.45	**82.57**

Table 1. The RE performance improvement by introducing a simple geometric restriction (on the FUNSD dataset).

1. Introduction

Visual information extraction (VIE) is a critical part in Document AI [3, 29, 47]. It has attracted more and more attention from both the academic and industrial community. VIE involves semantic entity recognition (SER, *a.k.a.* entity labeling) and relation extraction (RE, *a.k.a.* entity linking) from visually-rich documents (VrDs) such as forms and receipts [3, 17, 22, 35, 39, 41, 45, 46]. Recent years have witnessed the great power of pre-trained multi-modal models [1, 7, 8, 12, 15, 20–22, 30, 38, 40, 41, 43] in VIE tasks, especially the SER task. Compared with SER, the RE task, which aims at predicting the relation between semantic entities in documents, has not been fully explored and remains a challenging problem [12, 22]. RE is essential to provide additional structural information closer to human comprehension of the VrDs [45]. It makes the open-layout information extraction possible, e.g., for open-layout key-value linking and form-like items grouping.

It is widely accepted that document layout understanding is crucial for VIE [1, 7, 8, 15, 21, 22, 30, 38, 40, 41, 43], especially for RE [12, 22]. The geometric relationships, a specific form for describing document layout, are important for document layout representations [22, 27, 31]. Most

*Both authors contributed equally to this work.
[1]https://github.com/AlibabaResearch/AdvancedLiterateMachinery

previous pre-trained models for VrDs learn layout representations *implicitly* by adding coordinates into the model inputs, combining the relative position encoding or supervising by alignment-related pre-training tasks like text-image alignment [15, 30, 43] and masked vision language modeling [1,7,12,15,21,22,22,40,41,43]. However, it is not guaranteed that the geometric layout information is well learned in these models. Taking the state-of-the-art model LayoutLMv3 as an example, we find it would make mistakes in certain relatively simple scenarios, where the geometric relations between entities are not complicated. As shown in Fig. 1, LayoutLMv3 seems to link two entities depending more on the semantics than the geometric layout. This indicates that its layout understanding is not sufficiently discriminative. To further verify our conjecture, we conduct an experiment by filtering the false positive relations using a simple geometric restriction (the linkings between entities should not point up beyond a certain distance), the precision would increase by a large margin (more than 4 points) while the recall is controlled unchanged, as detailed in Tab. 1. This experiment proves that LayoutLMv3 does not fully exploit the useful geometric relationship information. Besides, most existing methods did not directly take the relation modeling into consideration in pre-training. They usually adopt token/segment-level classification or regression, which might underperform on downstream tasks related to relation modeling. Therefore, it is necessary to learn a better layout representation for document pre-trained models by modeling the geometric relationships between entities *explicitly* during pre-training.

During RE fine-tuning, previous works usually learn a task head like a single linear or bilinear layer [12, 22] from scratch. On the one hand, since the higher-level pair relationship features, which are beyond the token or text-segment features in documents, are complex, we argue that a single linear or bilinear layer is not always adequate to make full use of the encoded features for RE. On the other hand, the RE task head initialized randomly is prone to overfitting with limited fine-tuning data. Since the pre-trained backbone has shown tremendous potential [4, 5], why not pre-train the task head in some way simultaneously? Several works [10, 14, 26] have proved that smaller *gap* between pre-training and fine-tuning leads to better performance for downstream tasks. Hence, there is still considerable room for the design and usage of the RE task head.

Based on the above observations, we establish a multi-modal pre-trained framework (termed as **GeoLayoutLM**) for VIE, in which a geometric pre-training strategy is designed to explicitly utilize the geometric relationships between text-segments, and elaborately-designed RE heads are introduced to mitigate the gap between pre-training and fine-tuning on the downstream relation extraction task. Specifically, three geometric relations are defined: the relation between two text-segments (**GeoPair**), that among multiple text-segment pairs (**GeoMPair**), and that among three text-segments (**GeoTriplet**). Correspondingly, three self-supervised pre-training tasks are proposed. GeoPair relation is modeled by the **D**irection and **D**istance **M**odeling (**DDM**) task in which GeoLayoutLM needs to tell the direction of a directed pair and identify whether a segment is the nearest to another one in the direction. Furthermore, we design a brand-new pre-training objective called **D**etection of **D**irection **E**xceptions (**DDE**) for GeoMPair, enabling our model to capture the common pattern of directions among segment pairs, enhance the pair feature representation and discover the detached ones. For GeoTriplet, we propose a **C**ollinearity **I**dentification of **T**riplet (**CIT**) task to identify whether three segments are collinear, which takes a step forward to the modeling of multi-segments relations. It is important for non-local layout feature learning especially in form-like documents. Additionally, novel relation heads are proposed to learn better relation features, which are pre-trained by the geometric pre-training tasks to absorb prior knowledge about geometry, thus mitigating the gap between pre-training and fine-tuning. Extensive experiments on five public benchmarks demonstrate the effectiveness of the proposed GeoLayoutLM.

Our contributions are summarized as follows:

1) This paper introduces three geometric relations in different levels and designs three brand-new geometric pre-training tasks correspondingly for learning the geometric layout representation explicitly. To the best of our knowledge, GeoLayoutLM is the first to explore the geometric relations of multi-pair and multi-segments in document pre-training.
2) Novel relation heads are proposed to benefit the relation modeling. Besides, the relation heads are pre-trained by the proposed geometric tasks and fine-tuned for RE, thus mitigating the object gap between pre-training and fine-tuning.
3) Experimental results on visual information extraction tasks including key-value linking as relation extraction, entity grouping as relation extraction, and semantic entity recognition show that the proposed GeoLayoutLM significantly outperforms previous state-of-the-arts with good interpretability. Moreover, our model has notable advantages in few-shot RE learning.

2. Related Works

Visual information extraction Visual information extraction (VIE) aims at extracting entities from visually-rich document images, typically including semantic entity recognition (SER) and relation extraction (RE) [12, 17, 22, 42]. Early works based on graph neural networks [18, 27, 31, 33,

34, 36, 44] learned node features of text and layout in the downstream VIE tasks directly. Recently, pre-training techniques have boosted the performance on document understanding. Various pre-training tasks are designed to learn better text/image features and their alignment for stronger multimodal document representation [1, 7, 8, 15, 21, 22, 25, 30, 38, 40, 41, 43]. Although they have achieved significant improvement on SER, RE remains largely underexplored and is also a challenging task [12, 16, 22, 45]. BROS [12] encoded the relative spatial positions of texts into BERT [4] to learn the layout representation better. In this paper, we focus on adopting pre-training to obtain better features.

Geometric information Geometric information is an important clue to represent the document layout. Liu *et al*. [27] utilized relative 2D positions in GNN. GraphNEMR [31] incorporated the 8-geometry neighbours and geometry distance information in document modeling for SER. SPADE [16] re-formulated the self-attention layer by introducing a relative spatial vector which is composed of relative coordinates, distance and angle embeddings. StrucText [22] proposed a Paired Boxes Direction task to model the geometric direction of text-segments in pre-training. However, these works only explored pair-level geometric relations. We expand geometric relation to more than two segments: the relations of multi-pairs and triplets are fully explored.

Pre-training / Fine-tuning Recent studies on pre-training also focused on alleviating the gap between the pre-training stage and the downstream fine-tuning stage [2, 9, 10, 13, 14, 26]. Hu *et al*. [14] identified and studied the training schema gap and the task knowledge gap, and converted the downstream ranking task into a pre-training schema. Prompt-based models were proposed to adapt to various scenarios by converting downstream tasks to proper prompts which are consistent with the schema in pre-training [26]. Inspired by these works, we pre-train our elaborately-designed relation heads using the geometric tasks to absorb geometric knowledge adequately and improve its generalization in relation representation from the large-scale pre-training data.

3. GeoLayoutLM

GeoLayoutLM is a multi-modal framework for VIE. Geometric information is explicitly encoded and utilized by the novel geometry-based pre-training tasks and the pre-training of the elaborately-designed relation heads. Additionally, an effective strategy for RE inference is introduced.

3.1. Model Architecture

3.1.1 Backbone

Inspired by the two-stream structure in METER [6] and SelfDoc [21], the backbone of GeoLayoutLM is composed of an independent vision module, a text-layout module, and interactive visual and text co-attention layers. As shown in

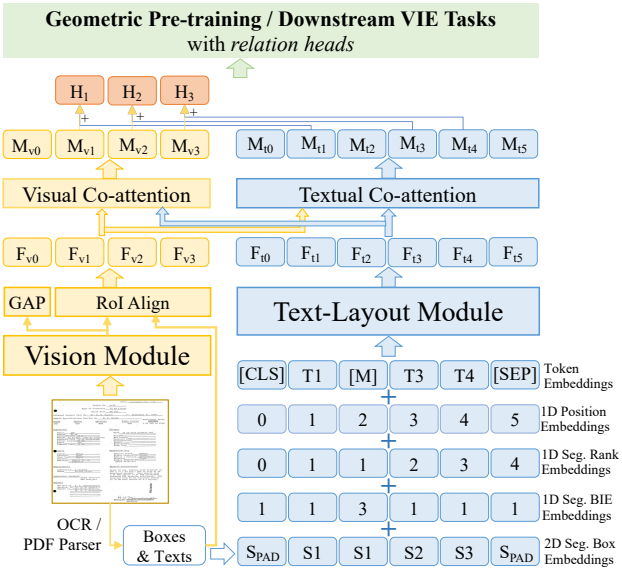


Figure 2. An overview of GeoLayoutLM.

Fig. 2, the vision module takes the document image as input, and the text-layout module is fed with layout-related text embeddings. Following LayoutLMv3 [15] and BiVLDoc [30], the text embeddings are the summation of 5 embeddings, including the token embeddings, 1D position embeddings, 1D segment rank embeddings, 1D segment BIE embeddings and 2D segment box embeddings. The output feature of the vision module is processed by the global average pooling [24] and RoI align [11] to compute the global visual feature F_{v0} and the n visual segment features $\{F_{vi}|i \in [1, n]\}$. Then the visual co-attention module takes $\{F_{vi}\}$ as query and $\{F_{ti}\}$ from the text-layout module as key and value for attention calculation, and outputs the fused visual features $\{M_{vi}\}$. The fused textual features $\{M_{ti}\}$ are calculated in a similar way. Finally, we add M_{vi} and the corresponding first token feature of the segment $M_{t,b(i)}$ to obtain the i-th segment feature H_i.

3.1.2 Relation Heads

The semantic entity recognition (SER) in VIE is usually modeled as a token classification problem. Learning a simple MLP classifier is effective for SER [15]. In the relation extraction (RE) of VIE, the final relation matrix was usually produced by a single linear or bilinear layer [12, 22]. Since the relationships of text-segments are relatively complex and related to each other, we argue that a simple linear or bilinear layer is not enough for relation modeling.

In this work, we propose two relation heads, including a Coarse Relation Prediction (CPR) head and a novel Relation Feature Enhancement (RFE) head, to enhance the relation feature representation for both relation pre-training and RE fine-tuning. The RFE head is a lightweight transformer [37]

consisting of a standard encoder layer, a modified decoder layer that discards the self-attention layer for computation efficiency, and a fully-connected layer followed by the sigmoid activation. As shown in Fig. 3, the text-segment features $\{H_i\}$ are fed into the CRP head (a bilinear layer) to predict a coarse relation matrix $r^{(0)}$. To build the relation features F_r, the segment feature pairs are passed to a pair feature extractor (linearly mapping the concatenated paired features). Then we select positive relation features F_r^+ based on $r^{(0)}$. Note that F_r^+ probably has some false positive relation features since $r^{(0)}$ is the coarse prediction. F_r^+ is then fed into the RFE encoder to capture the internal pattern of the true positive relations in each document sample, which is based on the assumption that most of the predicted positive pairs in $r^{(0)}$ are true. All the relation features F_r and the memory from the RFE encoder are fed into the RFE decoder to compute the final relation matrix $r^{(1)}$.

3.2. Pre-training

GeoLayoutLM is pre-trained with four self-supervised tasks simultaneously. To learn multimodal contextual-aware text representations, the widely-used Masked Visual-Language Model (MVLM) [15, 40, 43] is adopted on both $\{F_{ti}\}$ and $\{M_{ti}\}$. Three proposed self-supervised geometric pre-training tasks are described in Sec. 3.2.2.

3.2.1 Geometric Relationship

To better represent document layout by geometric information, three geometric relationships are introduced, which are **GeoPair**, **GeoMPair** and **GeoTriplet**. The relation between two text-segments (a pair) is denoted as **GeoPair**, which is also considered in previous works [22, 27, 31] to model the relative layout information between two text-segments. We further extend GeoPair to **GeoMPair** that is the relation among multiple segment pairs, to explore the relation of relations. Like the relation of three points in geometry, **GeoTriplet** is also devised, which is the relation among three text-segments.

3.2.2 Geometric Pre-training

To make our model understand the geometric relationships and achieve good layout representations, we propose three geometry-related self-supervised pre-training tasks to model GeoPair, GeoMPair, and GeoTriplet respectively. The input of these tasks are text-segments features $\{B_i\}$ which can be either of the five features: $\{H_i\}, \{M_{vi}\}, \{F_{vi}\}, \{M_{t,b(i)}\}, \{F_{t,b(i)}\}$, where $b(i)$ is the index of the first token of the i-th segment.

Direction and Distance Modeling for GeoPair To better understand the relative position relationship of two text-segments, as shown in Fig. 4(a), the Direction and Distance

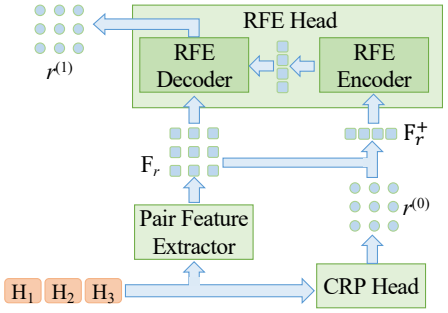


Figure 3. Relation heads.

Modeling (DDM) is proposed, in which both the direction and distance are measured.

We consider 9 directions, including 8 neighbor ones [31] and the overlapping. Hence, the direction modeling is exactly a 9-direction classification problem:

$$P_{ij}^{direct} = \text{Softmax}(\text{Linear}([B_i, B_j])) \tag{1}$$

where P_{ij}^{direct} is the predicted direction probability, $[\cdot]$ is the concatenation operation.

The distance between two segments is defined as the minimum distance between the two bounding boxes [31]. The distance is modeled as a binary classification problem that is to identify whether the j-th text-segment is the nearest to the i-th one in their direction. There is at most 1 nearest segment judged by distance in each of the 8 neighbor directions. A bilinear layer is applied here:

$$P_{ij}^{dist} = \text{Sigmoid}(\text{Bilinear}(B_i, B_j)) \tag{2}$$

where P_{ij}^{dist} is the probability of the nearest pair identification. Note that the operation in Eq. (2) shares the same process of CRP, which achieves the goals of pre-training the CRP head.

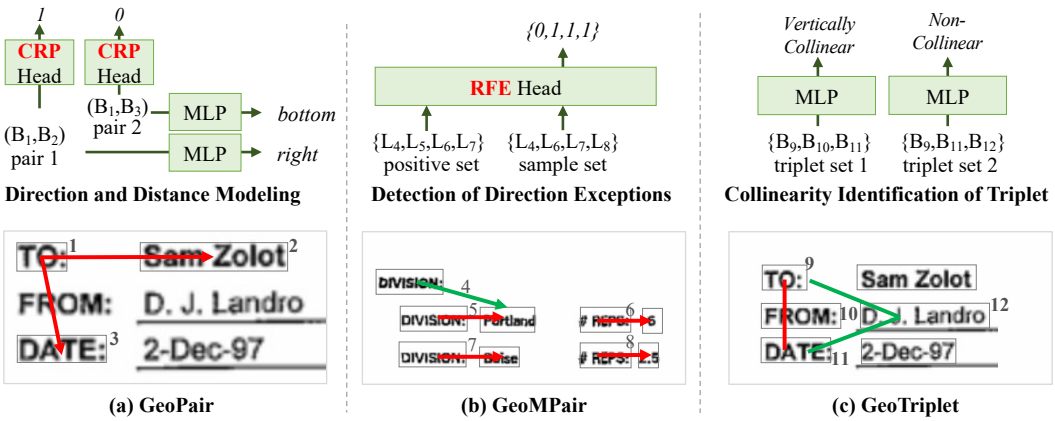
The loss function $\mathcal{L}_{DDM}$ of DDM task is defined as:

$$\begin{aligned}\mathcal{L}_{DDM} = &\text{CrossEntropy}(P^{direct}, Y^{direct}) \\ &+ \text{BCELoss}(P^{dist}, Y^{dist})\end{aligned} \tag{3}$$

where Y^{direct} and Y^{dist} are labels for direction and distance modeling.

Detection of Direction Exceptions for GeoMPair The relationships within a certain document area usually have some common geometric attributes. As shown in Fig. 4(b), the directions of key-value pairs are the same (arrows in red) in the wireless form area. The link in green can be easily judged as false due to its exceptional direction. Motivated by this, the Detection of Direction Exceptions (DDE) task is proposed to model GeoMPair for the non-local layout understanding in documents.

The DDE task is to discriminate segment pairs whether their directions are exceptional in a sample set S. A direction is regarded as an exception if the pairs with the direction have a minor ratio in the given positive set S_p. For



(a) GeoPair (b) GeoMPair (c) GeoTriplet

Figure 4. Geometric pre-training.

example, given a positive set in which more than 60% of pairs have the same direction *Right*, and a sample set, we label the right pairs in the sample set as 1, while the pairs of other directions as 0 (exception), as shown in Fig. 4(b). The pair feature L_i is built by linearly projecting the concatenated segment features. Then, the positive set and the sample set are fed into the proposed RFE head to predict the discrimination probabilities of the sample set. The binary cross-entropy loss is applied.

$$P^{DDE} = \text{RFE}(S_p, S) \tag{4}$$

$$\mathcal{L}_{DDE} = \text{BCELoss}(P^{DDE}, Y^{DDE}) \tag{5}$$

Collinearity Identification of Triplet for GeoTriplet The geometric alignment of segments is an important expression of document layout, which is meaningful and involves the relation of multiple segments. Like the collinear attribute of three points, we define that of three segments. Given three text-segments B_i, B_j and B_k, if the direction from B_i to B_j, B_j to B_k and B_i to B_k are the same or antiphase, they are collinear; otherwise non-collinear. The collinear cases can be further divided into four classes: horizontal line, vertical line, forward slash and backslash. As shown in Fig. 4(c), the left-aligned segments with the same entity tag are vertically collinear. Correspondingly, a pre-training task called Collinearity Identification of Triplet (CIT) is proposed. The triplet feature is the summation of three segment features since the collinear attribute is undirected. Subsequently, the 5-classification is made for CIT:

$$P^{CIT}_{ijk} = \text{Softmax}(\text{Linear}(B_i + B_j + B_k)) \tag{6}$$

$$\mathcal{L}_{CIT} = \text{CrossEntropy}(P^{CIT}, Y^{CIT}) \tag{7}$$

The full pre-training objective of GeoLayoutLM is:

$$\mathcal{L}_{pt} = \mathcal{L}_{MVLM} + \mathcal{L}_{DDM} + \mathcal{L}_{DDE} + \mathcal{L}_{CIT} \tag{8}$$

3.3. Fine-tuning and Inference

During fine-tuning, the relation heads are initialized with the pre-trained parameters, which mitigates the gap between pre-training and fine-tuning. The cross-entropy and binary cross-entropy function are utilized in SER loss $\mathcal{L}_{\text{SER}}$ and RE losses $\{\mathcal{L}_{\text{RE},i} | i = 0, 1\}$ (corresponds to $r^{(0)}$ and $r^{(1)}$) respectively. They are optimized together:

$$\mathcal{L}_{ft} = \mathcal{L}_{\text{SER}} + \sum_{i=0}^{1} \mathcal{L}_{\text{RE},i} \tag{9}$$

In the RE task, for a relation pair $B_j \rightarrow B_i$, B_j is called the father node, and B_i is the son node. $r^{(1)}_{i,j}$ stands for the probability that B_j is the father of B_i. The final relation output $\mathbb{R}$ was usually defined as: $\mathbb{R}_{ij} = \mathbb{1}\left(r^{(1)}_{i,j} > 0.5\right)$, where $\mathbb{1}(\cdot)$ is the indicator function. Optionally, we propose to impose the Restriction on the Selection of Fathers (RSF) for each son during the inference if some segments have several father nodes. Specifically, the j-th segment is regarded as a father node of the i-th one only if $r^{(1)}_{i,j} > 0.5$ and $r^{(1)}_{i,j}$ is close to the maximum probability:

$$\mathbb{R}_{ij} = \mathbb{1}\left(r^{(1)}_{i,j} > 0.5\right) \times \mathbb{1}\left(\max_k r^{(1)}_{i,k} < r^{(1)}_{i,j} + \tau\right) \tag{10}$$

where τ is the margin between the probabilities.

We suggest an additional variance loss for RSF especially. Since the probabilities of father nodes are expected to be as close as possible, the variance of them should be as small as possible. During fine-tuning, the variance loss is exerted on the son nodes that have more than 1 father node.

4. Experiments

4.1. Implementation Details

The backbone detail is described in Sec. 3.1.1. The vision module is composed of a ConvNeXt [28] and a multi-scale FPN [23]. BROS [12] is used as our text-layout module. The visual and textual co-attention modules are both equipped with a transformer decoder layer.

The document images are resized to 768×768. The embedding size and feed-forward size of the co-attention

module are 1024 and 4096 respectively. In the RFE head, the relation feature size and feed-forward size are both set to 1024. The number of attention heads is 2. τ is $1e-3$.

Pre-training Details. Following the LayoutLM series [15, 40, 43], we pre-train our model on the IIT-CDIP Test Collection 1.0 [19], which consists of around 11 million document images. However, the original line-level OCR annotation in the dataset will lead to the monotony of paired box directions (only top and bottom exist in most documents), which is extremely harmful to geometric pre-training. To break the imbalance in the distribution of the geometric relationship, we modify the OCR annotation by a Poisson Line Segmentation algorithm for each document with a probability of 90%. Algorithm 1 lists the procedure of splitting a line.

Algorithm 1: Poisson Line Segmentation

Input: An original line from OCR annotation L
Output: The processed line(s) L'

Get the number of words in $L_i \rightarrow N_w$;
$p_l = (1 - 1/(N_w - 0.5))$; `// split probability`
if $N_w < 2 || \text{rand}() > p_l$ **then**
  $L' \leftarrow L$
else
  $N_s = \textbf{poisson}(\lambda = \min(N_w/3, 7))$;
  Split L into N_s segments $\rightarrow L'$;
end

The AdamW optimizer is applied for pre-training, with the initial learning rate of 1e-5 and a linear decay learning rate scheduler. We use a batch size of 224 to train GeoLayoutLM for 2 epochs on the IIT-CDIP dataset. The maximum sequence length is set to 512. The maximum number of text-segments is set to 256. Following [15, 40, 43], we mask 15% text tokens in which 80% are replaced by the [MASK] token, 10% are replaced by a random token, and 10% keeps unchanged. In DDM, 16 text-segments are randomly sampled. Then, for each of them, we randomly sample 32 different text-segments to build paired text-segments. In DDE, 40 segment pairs are randomly sampled in the document. In CIT, 16 triple text-segments are randomly sampled.

Fine-tuning Details. Following the LayoutLM series [8, 15, 43], the SER task is regarded as a sequence labeling problem aiming to tag each word with a label. For the RE task, to conduct fair comparisons with previous methods (e.g., BROS [12]), the ground truth entity labels are used. We evaluate GeoLayoutLM on two popular benchmark datasets with five subtasks. **FUNSD** [17] is a scanned document dataset for form understanding. It has 149 training samples and 50 test samples with multifarious layouts. We focus on both the semantic entity recognition (a.k.a. entity labeling) and the relation extraction (a.k.a. entity linking) tasks. **CORD** [32] is a camera-captured receipt dataset for information extraction. It contains 800 training, 100 validation and 100 test images. In CORD, three subtasks are evaluated including semantic entity recognition (SER), relation extraction as entity grouping (REaGRP) and relation extraction as key-value linking (REaKV). We fine-tune our GeoLayoutLM for 200 epochs in FUNSD and 100 epochs in CORD with the batch size of 6. The learning rate is initially set to $2e-5$.

4.2. Comparison with the SOTAs

We compare our results with the previous state-of-the-arts. As shown in Tab. 2, our GeoLayoutLM obtains the best F1 score in both semantic entity recognition (SER) and relation extraction (RE).

For the FUNSD SER task, GeoLayoutLM and LayoutLMv3 both significantly surpass other models. Besides, the SER results on FUNSD and CORD also suggest that the geometric pre-training does the SER slightly more favorable than the popular text-image alignment. For the RE task, GeoLayoutLM significantly outperforms the previous state-of-the-art by 9.1% on FUNSD, and reaches or nearly reaches the perfect performance in CORD. It demonstrates the great superiority of our model in extracting relations. Even if we only fine-tune for 100 epochs, we still achieve (SER: 92.24%, RE: 88.80%) on FUNSD.

GeoLayoutLM backbone is slightly heavy due to the two-tower encoder. Our vision module is flexible and can be replaced by others. LayoutLMv3 has a coupling feature encoder for visual patches and text, which contributes to fewer parameters. The relation head we used in LayoutLMv3 is the same as the CPR head in GeoLayoutLM (1M Params). The proposed RFE head (14M) only constitutes 3.5% of the total parameters. On one Nvidia V100 GPU, the average inference time of GeoLayoutLM is 80.17ms, which is nearly the same as that of LayoutLMv3 (79.69ms).

4.3. Ablation Study

To better understand the effectiveness of geometric pre-training, the design of the RE heads and the RSF strategy in GeoLayoutLM, we perform plentiful ablation studies.

Impact of Geometric Pre-training. To figure out how each pre-training task influences the information extraction result, we pre-train our model using different combinations of the geometric tasks while remaining the MLM task. To be efficient, only 10% of the original pre-training data is used to train the model for 1 epoch. The results meet our expectations completely, as shown in Tab. 3. By comparing #0 with #1x, we observe that the performance of SER and RE will be improved if either of the three geometric tasks is exerted paralleled to the MLM task. For SER, GeoPair contributes the most while GeoMPair does the least. For RE, GeoMPair contributes the most while GeoTriplet does the least, which may be owing to the RE head that is directly

Method	#Params	FUNSD		CORD		
		SER	RE	SER	REaKV	REaGRP
BERT$_{LARGE}$ [4]	340M	65.63	29.11	90.25	-	-
LayoutLM$_{LARGE}$ [40]	343M	78.95	42.83	94.93	-	-
StrucTexT [22]	107M	83.09	44.10	-	-	-
SERA [45]	-	-	65.96	-	-	-
LayoutLMv2$_{LARGE}$ [43]	426M	84.20	70.57	96.01	-	97.29
BROS$_{LARGE}$ [12]	340M	84.52	77.01	97.28	-	97.40
LayoutLMv3$_{LARGE}$ [15]	357M	92.08	80.35†	97.46	99.64†	98.28†
GeoLayoutLM	399M	**92.86**	**89.45**	**97.97**	**100.00**	**99.45**

Table 2. Comparison with existing models that explore both SER & RE. The F1 score followed by † means it is re-implemented by us.

#	GeoPair	GeoMPair	GeoTriplet	SER	RE
0				83.39	74.91
1a	✓			91.80	82.23
1b		✓		88.67	82.56
1c			✓	90.78	78.90
2a	✓	✓		91.86	85.22
2b	✓		✓	91.90	82.37
2c		✓	✓	91.39	84.97
3	✓	✓	✓	92.17	85.32

Table 3. Ablation study on the geometric pre-training task in FUNSD. The first column labels the experiment settings.

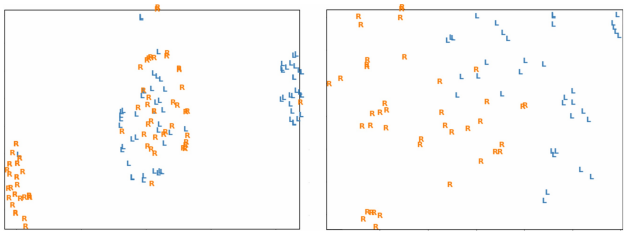
	Entropy↓	Cross Entropy↓	Acc.↑
LayoutLMv3	1.1423	0.9986	0.6319
GeoLayoutLM	**0.7633**	**0.5884**	**0.8223**

Table 4. Experiments on the geometric layout understandings. The entropy of direction prediction reveals the information maintained in the backbone. The lower the Entropy and the Cross Entropy are, the more layout information the model maintains.

pre-trained in GeoPair and GeoMPair. By comparing #2x with #1x and #3x with #2x, we find that it is always better to pre-train with more geometric tasks, which indicates that the tasks are complementary.

To be interpretable, we also investigate how much information of geometric relationship is kept after an example is encoded for the downstream information extraction task. To this end, we exert a linear classifier onto the backbone of the model fine-tuned on FUNSD (GeoLayoutLM VS LayoutLMv3), and *only train the classifier* on the re-processed FUNSD dataset with pair direction labels (9-direction classification), to squeeze the direction information that is measured by the classification entropy, cross-entropy and the accuracy. As shown in Tab. 4, GeoLayoutLM has a lower entropy and cross-entropy, and a higher accuracy, indicating that it retains much more information about geometric relations in the downstream tasks.

To further understand the geometric layout information,



(a) LayoutLMv3 (b) GeoLayoutLM

Figure 5. Comparison of the left (L) and right (R) relation features. For LayoutLMv3 and GeoLayoutLM, 2D layout positions remain unchanged and the input text tokens are set to [UNK].

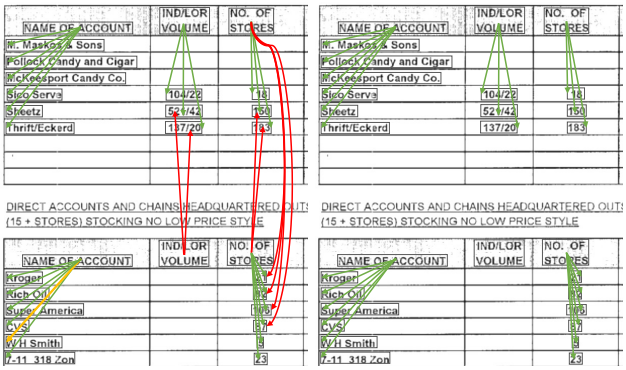
we conduct an embedding visualization of the left-right direction. As shown in Fig. 5, GeoLayoutLM has stronger distinguishable embeddings in left and right relationships that are important for document layout representation.

A case study is given in Fig. 6. Most of the false positive relation links predicted by LayoutLMv3 violate the geometric layout obviously. It depends on the semantic information excessively and ignores the layout more or less. For example, the entity starting with "No." is linked to the number entity regardless of the geometric relationship between them. In contrast, GeoLayoutLM successfully predicted all links with a good recall.

Although the rule-based geometric constraint can bring some improvement (Tab. 1), it still fall behind GeoLayoutLM because it: (1) relies on hard-coded thresholds, which limit its adaptability and generalization when handling documents of different formats and layouts; (2) is able to prune false linkings, but cannot recover missed ones.

Effects of the Relation Heads. There are two important points in our RE task heads: the novel RFE head and its pre-training. We study the impact of them for the RE task. The coarse relation prediction (CRP) head is always kept. Besides, we do not use the RSF strategy to be elegant.

As shown in Tab. 5, a bare CRP head not initialized by the pre-trained parameters (w/o Pt) achieves an 82.2% F1 score owing to the strong geometry-aware backbone. Once it is initialized by pre-training (Pt), an improvement of 2.7% F1 score is obtained. By adding the RFE head (w/o Pt),



(a) LayoutLMv3 (b) GeoLayoutLM

Figure 6. RE case study. The arrows in green, red and orange denote true positive, false positive and false negative (missed) relations respectively. Best viewed by zooming up.

CRP head		RFE head		F1
w/o Pt	Pt	w/o Pt	Pt	
✓				82.2
	✓			84.9
✓		✓		84.0
	✓	✓		84.0
✓			✓	86.2
	✓		✓	**86.9**

Table 5. Ablation study on the CRP and RFE head in FUNSD RE task. "w/o Pt" and "Pt" mean that the head is not pre-trained and pre-trained respectively. The RSF strategy is not used.

the version of CRP (w/o Pt) becomes stronger while that of CRP (Pt) even degrades a little bit. We argue that the RFE head introduces more parameters, which causes overfitting despite its superiority in relation modeling. Thus it is necessary to pre-train the RFE head. We also find that the pre-training of the RFE head is more important than that of the CRP head. By making full use of the two points, GeoLayoutLM obtains the best RE performance.

Effects of RSF. The RSF strategy is non-trivial in the fine-tuning and inference stage. It contains two parts: the post-process for inference and the variance loss in fine-tuning.

Tab. 6 gives a clear view. By the post-processing, the precision of our method is improved dramatically with a little sacrifice of recall. A bare variance loss without the post-processing does nothing to the performance since it is designed for the post-processing only. We obtain the best F1 score when using both of them.

4.4. Few-shot RE Learning

In real scenarios, the acquirement of the training data for document information extraction is a bottleneck due to the expensive and boring annotation work. So it is necessary to learn from only a few document samples.

To explore the ability of few-shot learning, we compare

postprocess	variance loss	Precision	Recall	F1
		85.26	90.15	87.64
✓		88.25	89.01	88.62
	✓	85.06	90.34	87.62
✓	✓	88.94	89.96	89.45

Table 6. Ablation study on the RSF strategy in FUNSD RE task.

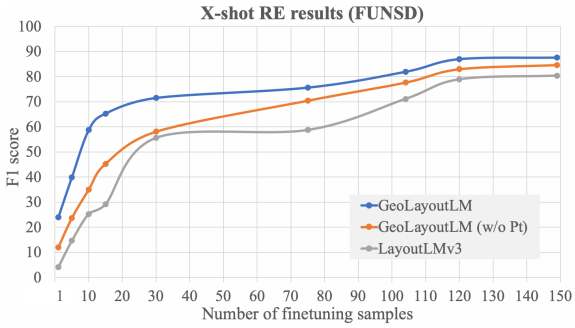


Figure 7. Comparison of few-shot learning in FUNSD RE task.

our GeoLayoutLM with another two models: a modified GeoLayoutLM whose heads are not initialized from pre-training (GeoLayoutLM*), and LayoutLMv3. We also disable the RSF strategy to make it clearer.

As shown in Fig. 7, GeoLayoutLM shows great superiority in this setting. GeoLayoutLM* outperforms LayoutLMv3 but is inferior to GeoLayoutLM all the time. It suggests that the geometric pre-training endows our model with a strong ability to extract entity relations, and also emphasizes the importance of RE head pre-training. Notably, our GeoLayoutLM achieves a slightly better performance (71.53%) using only 30 samples than LayoutLMv3 does (71.07%) using 104 samples. The performance gap is very large when only few fine-tuning samples are available.

5. Conclusion

In this paper, we propose GeoLayoutLM, a geometric pre-training framework for VIE. Three geometric relations in different levels are defined: GeoPair, GeoMPair and GeoTriplet. Correspondingly, three specially designed pre-training objectives are introduced to model geometric relations explicitly. Additionally, the relation heads are elaborately designed to enhance the relation feature representation, which are pre-trained by the geometric pre-training, thus mitigating the gap between pre-training and fine-tuning. Experimental results on VIE have illustrated the effectiveness of GeoLayoutLM in both SER and RE tasks. In the future, we will explore more effective geometric pre-training tasks, and apply our method to more tasks of visually-rich document understanding.

References

[1] Srikar Appalaraju, Bhavan Jasani, and Bhargava Urala Kota. DocFormer: End-to-end transformer for document understanding. In *ICCV*, pages 4171–4186, 2021. 1, 2, 3

[2] Alexandra Chronopoulou, Christos Baziotis, and Alexandros Potamianos. An embarrassingly simple approach for transfer learning from pretrained language models. *NAACL*, 2019. 3

[3] Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. Document ai: Benchmarks, models and applications. *arXiv preprint arXiv:2111.08609*, 2021. 1

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 3, 7

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[6] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *CVPR*, pages 18166–18176, 2022. 3

[7] Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Nikolaos Barmpalios, Rajiv Jain, Ani Nenkova, and Tong Sun. Unified pretraining framework for document understanding. In *NeurIPS*, 2021. 1, 2, 3

[8] Zhangxuan Gu, Changhua Meng, Ke Wang, Jun Lan, Weiqiang Wang, Ming Gu, and Liqing Zhang. Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding. In *CVPR*, pages 4583–4592, 2022. 1, 3, 6

[9] Suchin Gururangan, Ana Marasovi'c, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: adapt language models to domains and tasks. *ACL*, 2020. 3

[10] Xueting Han, Zhenhuan Huang, Bang An, and Jing Bai. Adaptive transfer learning on graph neural networks. In *SIGKDD*, 2021. 2, 3

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 3

[12] Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. In *AAAI*, 2022. 1, 2, 3, 5, 6, 7

[13] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *ACL*, 2018. 3

[14] Xiaomeng Hu, Shi Yu, Chenyan Xiong, Zhenghao Liu, Zhiyuan Liu, and Ge Yu. P ^{3} ranker: Mitigating the gaps between pre-training and ranking fine-tuning with prompt-based learning and pre-finetuning. *SIGIR*, 2022. 2, 3

[15] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *ACM Multimedia*, 2022. 1, 2, 3, 4, 6, 7

[16] Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang, and Minjoon Seo. Spatial dependency parsing for semi-structured document information extraction. *ACL findings*, 2021. 3

[17] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *ICDARW*, volume 2, pages 1–6, 2019. 1, 2, 6

[18] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 2

[19] David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. Building a test collection for complex document information processing. In *SIGIR*, 2006. 6

[20] Chenliang Li, Bin Bi, and Ming Yan. StructuralLM: Structural pre-training for form understanding. In *ACL*, 2021. 1

[21] Peizhao Li, Jiuxiang Gu, and Jason Kuen. SelfDoc: Self-supervised document representation learning. In *CVPR*, pages 5652–5660, 2021. 1, 2, 3

[22] Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. Structext: Structured text understanding with multi-modal transformers. In *ACM Multimedia*, pages 1912–1920, 2021. 1, 2, 3, 4, 7

[23] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *AAAI*, volume 34, pages 11474–11481, 2020. 5

[24] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In *ICLR*, 2014. 3

[25] Weihong Lin, Qifang Gao, Lei Sun, Zhuoyao Zhong, Kai Hu, Qin Ren, and Qiang Huo. Vibertgrid: a jointly trained multi-modal 2d document representation for key information extraction from documents. In *ICDAR*, pages 548–563. Springer, 2021. 3

[26] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, sep 2022. Just Accepted. 2, 3

[27] Xiaojing Liu, Feiyu Gao, and Qiong Zhang. Graph convolution for multimodal information extraction from visually rich documents. In *NAACL-HLT (1)*, pages 32–39, 2019. 1, 2, 3, 4

[28] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022. 5

[29] Shangbang Long, Xin He, and Cong Yao. Scene text detection and recognition: The deep learning era. *International Journal of Computer Vision*, 129:161–184, 2018. 1

[30] Chuwei Luo, Guozhi Tang, Qi Zheng, Cong Yao, Lianwen Jin, Chenliang Li, Yang Xue, and Luo Si. Bi-vldoc: Bidirectional vision-language modeling for visually-rich document understanding. *arXiv preprint arXiv:2206.13155*, 2022. 1, 2, 3

[31] Chuwei Luo, Yongpan Wang, Qi Zheng, Liangchen Li, Feiyu Gao, and Shiyu Zhang. Merge and recognize: a geometry and 2d context aware graph model for named entity recognition from visual documents. In *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*, pages 24–34, 2020. 1, 2, 3, 4

[32] Seunghyun Park, Seung Shin, Bado Lee, and Junyeop Lee. CORD: A consolidated receipt dataset for post-ocr parsing. In *Document Intelligence Workshop at NeurIPS*, 2019. 6

[33] Yujie Qian, Enrico Santus, and Zhijing Jin. GraphIE: A graph-based framework for information extraction. In *NAACL*, pages 751–761, 2019. 2

[34] Yujie Qian, Enrico Santus, Zhijing Jin, Jiang Guo, and Regina Barzilay. Graphie: A graph-based framework for information extraction. In *NAACL*, 2019. 2

[35] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:2298–2304, 2015. 1

[36] Guozhi Tang, Lele Xie, Lianwen Jin, and Wang. MatchVIE: Exploiting match relevancy between entities for visual information extraction. In *IJCAI*, 2021. 2

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[38] Jiapeng Wang, Lianwen Jin, and Kai Ding. Lilt: A simple yet effective language-independent layout transformer for structured document understanding. In *ACL*, 2022. 1, 3

[39] Peng Wang, Cheng Da, and Cong Yao. Multi-granularity prediction for scene text recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 339–355. Springer, 2022. 1

[40] Yiheng Xu, Minghao Li, Lei Cui, and Shaohan Huang. LayoutLM: Pre-training of text and layout for document image understanding. In *KDD*, pages 1192–1200, 2020. 1, 2, 3, 4, 6, 7

[41] Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. *arXiv preprint arXiv:2104.08836*, 2021. 1, 2, 3

[42] Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. Xfund: A benchmark dataset for multilingual visually rich form understanding. In *ACL Findings*, pages 3214–3224, 2022. 2

[43] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *ACL*, 2021. 1, 2, 3, 4, 6, 7

[44] Wenwen Yu, Ning Lu, Xianbiao Qi, Ping Gong, and Rong Xiao. Pick: Processing key information extraction from documents using improved graph learning-convolutional networks. In *ICPR*, 2020. 2

[45] Yue Zhang, Bo Zhang, Rui Wang, Junjie Cao, Chen Li, and Zuyi Bao. Entity relation extraction as dependency parsing in visually rich documents. *EMNLP*, 2021. 1, 3, 7

[46] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: An efficient and accurate scene text detector. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2642–2651, 2017. 1

[47] Yingying Zhu, Cong Yao, and Xiang Bai. Scene text detection and recognition: recent advances and future trends. *Frontiers of Computer Science*, 10:19–36, 2016. 1