



ReLaText: Exploiting visual relationships for arbitrary-shaped scene text detection with graph convolutional networks



Chixiang Ma^{a,b,1,*}, Lei Sun^b, Zhuoyao Zhong^b, Qiang Huo^b

^a Department of EELS, University of Science and Technology of China, Hefei, 230026, China

^b Microsoft Research Asia, Beijing, 100080, China

ARTICLE INFO

Article history:

Received 15 March 2020

Revised 26 June 2020

Accepted 24 September 2020

Available online 25 September 2020

Keywords:

Arbitrary-Shaped text detection

Graph convolutional network

Link prediction

Visual relationship detection

ABSTRACT

We introduce a new arbitrary-shaped text detection approach named ReLaText by formulating text detection as a visual relationship detection problem. To demonstrate the effectiveness of this new formulation, we start from using a “link” relationship to address the challenging text-line grouping problem firstly. The key idea is to decompose text detection into two subproblems, namely detection of text primitives and prediction of link relationships between nearby text primitive pairs. Specifically, an anchor-free region proposal network based text detector is first used to detect text primitives of different scales from different feature maps of a feature pyramid network, from which a text primitive graph is constructed by linking each pair of nearby text primitives detected from a same feature map with an edge. Then, a Graph Convolutional Network (GCN) based link relationship prediction module is used to prune wrongly-linked edges in the text primitive graph to generate a number of disjoint subgraphs, each representing a detected text instance. As GCN can effectively leverage context information to improve link prediction accuracy, our GCN based text-line grouping approach can achieve better text detection accuracy than previous text-line grouping methods, especially when dealing with text instances with large inter-character or very small inter-line spacing. Consequently, the proposed ReLaText achieves state-of-the-art performance on five public text detection benchmarks, namely RCTW-17, MSRA-TD500, Total-Text, CTW1500 and DAST1500.

© 2020 Elsevier Ltd. All rights reserved.

1. Introduction

Scene text detection has received considerable attention from computer vision and document analysis communities recently [1–5], due to its important role in many content-based visual intelligent applications like image retrieval, autonomous driving and OCR translation. Unlike traditional OCR techniques that only deal with texts in scanned document images, scene text detection tries to detect arbitrary-shaped texts from complex natural scene images, in which texts usually occur on street nameplates, store signs, restaurant menus, product packages, advertising posters, etc. Due to high variations of text font, color, scale, orientation and language, extremely complex backgrounds, as well as various distortions and artifacts caused by image capturing like non-uniform

illumination, low contrast, low resolution, and occlusion, scene text detection is still an unsolved problem.

In recent years, convolutional neural network (CNN) based scene text detection approaches have made great progress and substantially outperformed traditional MSER or SWT based text detection methods (e.g., [6–8]) in terms of both accuracy and capability. As earlier scene text benchmark datasets like ICDAR'13, '15, '17² do not contain curved texts, most earlier methods [1–5,9–17] simply assume that texts are arranged in a straight line manner and adapt CNN-based object detection and segmentation frameworks, like Faster R-CNN [18], SSD [19], YOLO [20], DenseBox [21] and FCN [22], to solve the text detection problem. Although these methods can deal with horizontal and multi-oriented text-lines well, they are incapable of detecting curved texts. Recently, with the emergence of two curved-text focused datasets, namely Total-Text [23] and CTW1500 [24], the research focus of text detection has shifted from horizontal or multi-oriented scene texts to more challenging curved or arbitrary-shaped scene texts. This has spawned lots of more effective approaches which can be roughly

* Corresponding author.

E-mail addresses: ma1996@mail.ustc.edu.cn (C. Ma), lsun@microsoft.com (L. Sun), zhuoyao.zhong@gmail.com (Z. Zhong), qianghuo@microsoft.com (Q. Huo).

¹ This work was done when Chixiang Ma was an intern in Speech Group, Microsoft Research Asia, Beijing, China.

² <https://rrc.cvc.uab.es/>.

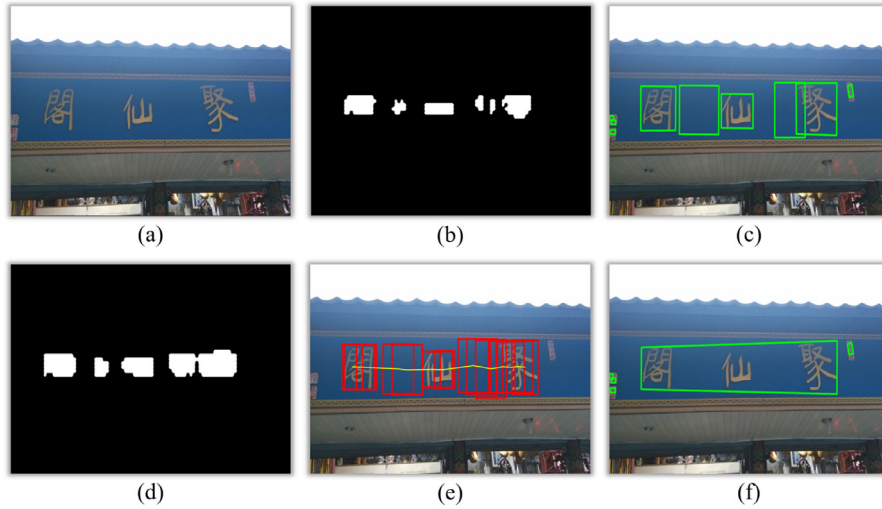


Fig. 1. (a) Original image from RCTW-17; (b) Binarized textness score map from SegLink [4]; (c) Detection results of SegLink; (d) Binarized textness score map from our method; (e) Grouped text primitives of our method based on predicted link relationships; (f) Detection results of our method.

classified into two categories: top-down approaches and bottom-up approaches. Top-down approaches usually use Region Proposal Network (RPN) [18] to generate coarse rectangular proposals first, then adopt a polygonal bounding box regression module [24,25] or an instance mask prediction module [26–29] to predict a tighter polygon-shaped bounding box for each curved text. It has been found that these approaches are not robust to nearby long curved text instances as in the DAST1500 dataset [30], because the rectangular proposals in this scenario are highly overlapped, which will cause some of them to be wrongly suppressed by the non-maximum suppression (NMS) algorithm and hurt the recall rate. Unlike top-down approaches, bottom-up approaches can get rid of the limitation of rectangular proposals, where the whole text instances are generated by grouping the detected candidate text segments (characters or parts of words/text-lines) or pixels. However, the difficulties lie in how to solve the text-line grouping problem robustly. Segment based methods have tried rule and character embedding based line grouping algorithms to solve this problem (e.g., [13,31,32]), but their results on curved text datasets are still worse than pixel-based methods. Pixel-based methods use local pixel connectivity to merge pixels on binarized textness score maps and leverage various auxiliary information (e.g., [4,33–35]) to improve pixel merging accuracy. Although achieving superior performance on benchmark datasets, these methods cannot detect text instances with large inter-character spacing robustly because pixels within large inter-character gaps tend to be misclassified as non-text (Fig. 1(b)) so that the text-line is over-segmented (Fig. 1(c)).

To address the above problems, we propose to solve the text detection problem from a new perspective by formulating text detection as a visual relationship detection problem instead of an object detection or segmentation problem. In the field of visual relationship detection, visual relationships are defined as $\langle \text{subject}, \text{predicate}, \text{object} \rangle$ triplets, where the “subject” is related to the “object” by the “predicate” relationship. Text detection can be easily formulated as a visual relationship detection problem, namely, the “subject” or “object” can be defined as a text primitive which represents a text segment or a whole word/text-line, and the “predicate” can be defined as some relationships between text primitive pairs. We argue that visual relationship detection is a more flexible problem formulation for text detection. Because, with this new formulation, some existing difficult problems in text detection like

text-line grouping and duplicate removal can be solved in a unified framework by formulating them as some specific relationship prediction problems between text primitive pairs. For example, for text-line grouping, a “link” relationship can be defined to indicate whether two text segments belong to a same word/text-line. To solve the duplicate removal problem, a “duplicate” relationship can be defined to determine whether two overlapped bounding boxes correspond to a same word/text-line. Moreover, the visual relationship detection framework can also be applied to other document understanding tasks like key-value pair extraction from documents [36] and table recognition [37]. With a same problem formulation, these systems can be seamlessly integrated together making it much easier to leverage multi-task learning to improve the accuracy for each task.

In this work, to demonstrate the effectiveness of this new formulation, we start from using a “link” relationship to address the challenging text-line grouping problem firstly. To this end, we decompose text detection into two key subproblems, namely detection of text primitives and prediction of link relationships between text primitive pairs. Based on the predicted link relationships, text primitives are grouped into text instances. In our conference paper [38], we have demonstrated that leveraging link relationships predicted by a relation network [39] to do text-line grouping can achieve better text detection accuracy than previous local pixel connectivity based text-line grouping methods. In this paper, we introduce further a new Graph Convolutional Network (GCN) based link prediction approach [40], which can leverage context information more effectively than relation network to improve the link prediction accuracy so that even text instances with large inter-character or very small inter-line spacing can be robustly detected by our text detector. Consequently, the proposed text detector, ReLaText, achieves state-of-the-art performance on five public text detection benchmarks, namely RCTW-17, MSRA-TD500, Total-Text, CTW1500 and DAST1500.

Although the preliminary results of leveraging link relationship prediction to solve the text-line grouping problem have been reported in our conference paper [38], we extend it in this paper significantly in the following aspects: (1) We adopt GCN to replace relation network to do link relationship prediction and achieve improved text detection accuracy; (2) More ablation studies are conducted to demonstrate the effectiveness of our GCN based link relationship prediction approach; (3) Related works on scene text

detection, visual relationship detection and graph convolutional network are reviewed more comprehensively; (4) Experimental results on two more public benchmark datasets, namely RCTW-17 and DAST1500, are presented to compare our approach with other state-of-the-art approaches.

2. Related work

2.1. Scene text detection

Before the deep learning era, there were only a few works paying attention to arbitrary-shaped text detection. Shivakumara et al. [41] proposed a quad tree based approach to detecting curved texts from videos. Fabrizio et al. [42] proposed to group extracted candidate text CCs into a graph, in which arbitrary-shaped text-lines were detected based on some regularity properties. As the performance of these methods heavily depends on heuristic rules or handcrafted features, they are not as robust as recent deep learning based approaches.

With the rapid development of deep learning, numerous CNN based text detection methods have been proposed and substantially outperformed traditional methods by a big margin in terms of both accuracy and capability. These methods can be roughly classified into two categories: top-down methods and bottom-up methods.

Top-down methods. Top-down methods typically treat text as a special kind of object, and directly adapt state-of-the-art top-down object detection or instance segmentation frameworks to solve the text detection problem. Jaderberg et al. [10] adopted R-CNN [43] for text detection first, but its performance was limited by the traditional region proposal generation methods [44]. Later, Zhong et al. [45], Liao et al. [2] and Gupta et al. [11] adopted Faster R-CNN, SSD and YOLO to detect horizontal texts respectively. To extend Faster R-CNN and SSD to multi-oriented text detection, Ma et al. [12] and Liu et al. [3] proposed to utilize rotated rectangular or quadrilateral anchors to hunt for inclined text proposals. Since directly predicting the vertex coordinates of quadrilateral bounding boxes suffers from a label confusion issue about the vertex order, Liu et al. [46] proposed to discretize the bounding box into key edges and learn the correct match-type with a multi-class classifier. Moreover, as the anchor mechanism used by Faster R-CNN and SSD is inflexible for text detection tasks, Zhou et al. [5] and He et al. [1] followed the “anchor-free” idea of DenseBox [21] and proposed to use an FCN [22] to directly output the pixel-wise textness scores and bounding boxes of the concerned text instances through all locations and scales of an image. Although more flexible, the capabilities of DenseBox-based one-stage text detectors are limited because they cannot detect long or large text instances effectively [5]. To address this issue, Zhong et al. [47] proposed to use DenseBox to replace the original anchor-based RPN in Faster R-CNN so that their Faster R-CNN based text detector could get rid of the limitations of anchor mechanism while preserving good accuracy for multi-oriented text detection. Another method [14] first generated candidate boxes by sampling and grouping the detected corner points of text bounding boxes, among which unreasonable boxes were eliminated by position sensitive segmentation scores. Since the rectangular or quadrilateral bounding boxes predicted by the above-mentioned text detectors cannot enclose curved texts tightly enough, these methods cannot detect curved texts effectively. To extend R-FCN [48] to curved text detection, Liu et al. [24] modified the bounding box regression module to predict a tighter polygonal bounding box with 14 points for each text proposal, which was further refined by a recurrent neural network to make the boundary more accurate. Wang et al. [25] argued that polygons of fixed 14 points were not precise enough for long curved text-lines, so they proposed to use a re-

current neural network to predict polygons of different numbers of points for texts of different shapes. Meanwhile, another category of methods [26–29] formulated text detection as an instance segmentation problem and borrowed existing top-down instance segmentation frameworks like Mask R-CNN [49] to predict a segmentation mask and optionally extra geometric attributes for the corresponding text instance in each positive proposal. Although these methods, especially the Mask R-CNN based ones [27,28], have achieved superior performance on most benchmark datasets like Total-Text and CTW1500, they are not robust to nearby long curved text instances. Tang et al. [30] introduced a new dense and arbitrary-shaped text detection dataset, i.e., DAST1500, which is mainly composed of commodity images, to demonstrate this. The main reason is that the rectangular proposals of nearby long curved text instances generated by existing top-down methods are highly overlapped, and some of them may be wrongly suppressed by the non-maximum suppression (NMS) algorithm so that the corresponding text instances cannot be detected correctly.

Bottom-up methods. Bottom-up methods generally follow a component-grouping paradigm, i.e., detect text components first and then group these components into text instances. Compared with top-down methods, bottom-up methods can get rid of the limitations of the region proposal generation module. Based on the granularity of text components, these methods can be further divided into two categories: pixel-level methods and segment-level methods.

1) Pixel-level: Pixel-based methods usually leverage semantic segmentation or instance segmentation frameworks to predict a pixel-level textness score map firstly, then use different methods to group text pixels into words/text-lines and calculate the corresponding bounding boxes. Zhang et al. [9] first used FCN to predict text blocks from which character candidates were extracted with MSER, then post-processing methods were used to generate text-lines. More recent works in this category directly used local pixel connectivity (e.g., 8-neighbourhood) to merge pixels on binarized textness score maps into CCs, each of which represents a word/text-line. In order to avoid merging nearby words/text-lines together or over-segmenting words/text-lines into pieces, these approaches tried to leverage other auxiliary information, e.g., link prediction [4,50], progressive scale expansion [51,52], text border prediction [33], text center line extraction [35,53], text center-border probability prediction [54], Markov clustering [55], direction field prediction [34], pixel embedding mapping [56] and character affinity estimation [57] to enhance pixel merging performance. Although these local pixel connectivity based line grouping approaches have achieved superior performance on benchmark datasets, we find that they tend to over segment text instances with large inter-character spacing into pieces, which is also mentioned in [4,34,35].

2) Segment-level: Segment-based methods detect text segments firstly, each of which contains a character or part of a word/text-line. The difficulties of these approaches also lie in how to robustly group the detected text segments into words/text-lines. Earlier works in this category like CTPN [13] and Word-sup [31] adopted rule-based methods to group the detected text segments into horizontal or multi-oriented text instances, which were not robust to curved texts. Recently, Liu et al. [32] proposed a character embedding based method to group detected characters into curved text-lines. However, their reported results on Total-Text are worse than pixel-level methods. Our proposed ReLaText is also a segment-level bottom-up approach, but we formulate text detection as a visual relationship detection problem and take advantage of the graph convolutional network to predict link relationships between text segments so that more robust text-line grouping can be achieved for arbitrary-shaped texts.

2.2. Visual relationship detection

Visual relationship detection has experienced a rapid development since some large scale datasets like VRD [58] were released. The goal of visual relationship detection is to detect objects along with predicting the relationships between object pairs from images. There exist a variety of visual relationships in the real world, including spatial relationships, attributive relationships, interactive relationships, etc [39]. Among them, the spatial relationships, which are used to describe relative locations between objects, have also been widely studied in the document analysis field. For example, in table structure recognition, identifying the horizontal and vertical adjacency relations between cells is a crucial step to reconstruct the cellular structures of tables [37]. In the mathematical expression recognition task, one of the most important problems is to predict the spatial relationships, like *horizontal*, *above*, *subscript*, between symbols [59].

A typical paradigm for visual relationship detection is composed of three modules, i.e., individual object detection, subject-object pair construction and relationship classification [58,60–65]. As context information is important for improving relationship classification accuracy, most previous methods tried to leverage wider context by simultaneously taking both “subject” and “object” proposals as well as their union as input to predict the “predicate” relationship [58,60–62]. Later works [63,64,66] proposed to use GCNs or its variants to further enhance the context information. Other than context information, the semantic relationships between objects and their corresponding predicates are also very important for improving accuracy [58,62,67]. To leverage this information, Lu et al. [58] leveraged language priors from semantic word embeddings to finetune the likelihood of a predicted relationship. Yu et al. [62] proposed to distill the internal and external linguistic knowledge into a deep neural network to regularize visual model learning. In this work, we propose a new GCN based link relationship prediction approach for improving link prediction accuracy.

2.3. Graph convolutional network

Although deep learning has revolutionized many machine learning tasks with data typically represented in the Euclidean space, there are a large number of real-world applications requiring to deal with non-Euclidean data, which is usually represented with arbitrarily structured graphs and imposes significant challenges on existing machine learning algorithms. Recently, many studies on extending deep learning approaches for graph data have emerged, and we refer readers to [68] for a comprehensive survey. Early works first attempted to use recursive neural network to process such graph-structured data [69]. Follow-up works raised an increasing interest in generalizing convolutions to the graph domain to decompose the complicated computational operations over graph-structured data into a series of local operations for each node at each time step. The advances in this direction can be divided into two categories: spectral-based approaches and spatial-based approaches. In general, spectral-based approaches [70,71] process graph data from the perspective of graph signal processing, and define the convolutional operation based on the spectral graph theory. On the contrary, spatial-based approaches [72–74] define the convolutional operation directly on the graph, only involving the spatially close neighbors. This series of works are the most related to this paper. One of the challenges of these approaches is to design the graph convolutional operator to work on nodes with different degrees. Duvenaud et al. [75] achieved it by learning an individual weight matrix for each node degree. Hamilton et al. [74] adopted sampling strategy to obtain a fixed number of neighbors for each node. Later, Kipf et al.

[40] simplified the spectral-based graph convolutions with a localized first-order approximation, which bridged the gap between spectral-based approaches and spatial-based approaches. More recently, the latest advances on GCNs include attentional weighted edges [76], dynamic edge convolution [77], etc.

3. Methodology

3.1. Overview

We propose to represent each text instance as a sparse directed subgraph, where nodes and edges represent text primitives and link relationships between nearby text primitive pairs, respectively. Here, each text primitive represents a text segment as in the work [13]. Based on this representation, our ReLaText is designed to comprise three key modules: 1) An anchor-free RPN based text detection module to detect text primitives from a feature pyramid generated by a Feature Pyramid Network (FPN) backbone [78]; 2) A text primitive graph construction module to group text primitives detected from different feature maps into different subgraphs respectively by linking each pair of nearby text primitives from a same feature map with an edge; 3) A GCN-based link relationship prediction module to prune wrongly-linked edges in all subgraphs to generate the final results. A brief pipeline of ReLaText is illustrated in Fig. 2, and the details of each module are described in the following subsections.

3.2. Text primitive detection

We adopt an Anchor-Free RPN (AF-RPN) method [47] to detect text primitives from a feature pyramid generated by FPN, which is built on top of ResNet-50 [79]. Here, the feature pyramid consists of three levels, i.e., P_2 , P_3 and P_4 , whose strides are 4, 8 and 16 pixels, respectively. All feature pyramid levels have $C = 256$ channels. The key idea of AF-RPN is to use different DenseBox-based detection modules [21] to detect text instances of different scales from different feature pyramid levels so that it is more robust to large text-scale variance. To define the scales of arbitrary-shaped text instances, we represent the boundary of each text instance by a polygon with a fixed number of anchor point pairs on its two long sides (Fig. 3(a-c)). Based on this, we define the scale of a text instance as the length of its shorter side, which is estimated by the average length of lines connecting anchor point pairs (light blue lines in Fig. 3(c)). In the training stage, each ground-truth polygon is only assigned to one feature pyramid level according to its scale and ignored by other feature pyramid levels. Then, we use three scale-specific detection modules to detect text primitives contained in small (4px–23px), medium (24px–48px) and large (>48px) text instances from P_2 , P_3 and P_4 , respectively.

Furthermore, we borrow the idea of border learning [33] to enhance the robustness of each detection module to nearby text-lines. Specifically, for each ground-truth polygon in a raw image, we shrink its short side by a scaling factor of 0.5 to create a corresponding text core region (red region in Fig. 3(d)). Meantime, we expand its short side by a scaling factor of 1.2 and define the region outside the text core region but inside the expanded polygon as a border region (green region in Fig. 3(d)). Regions outside all expanded polygons are background regions. During training, each pixel on each pyramid level can be mapped back to a sliding point [47] in the raw image. For a pixel on a pyramid level, if its corresponding sliding point locates in the text core or border region of a ground-truth polygon assigned to this pyramid level, it is labeled as “text” or “border”, respectively. Pixels whose corresponding sliding points locate in ground-truth polygons assigned to other pyramid levels are ignored. The remaining pixels are labeled as “non-text”. For each “text” pixel, we use the algorithm

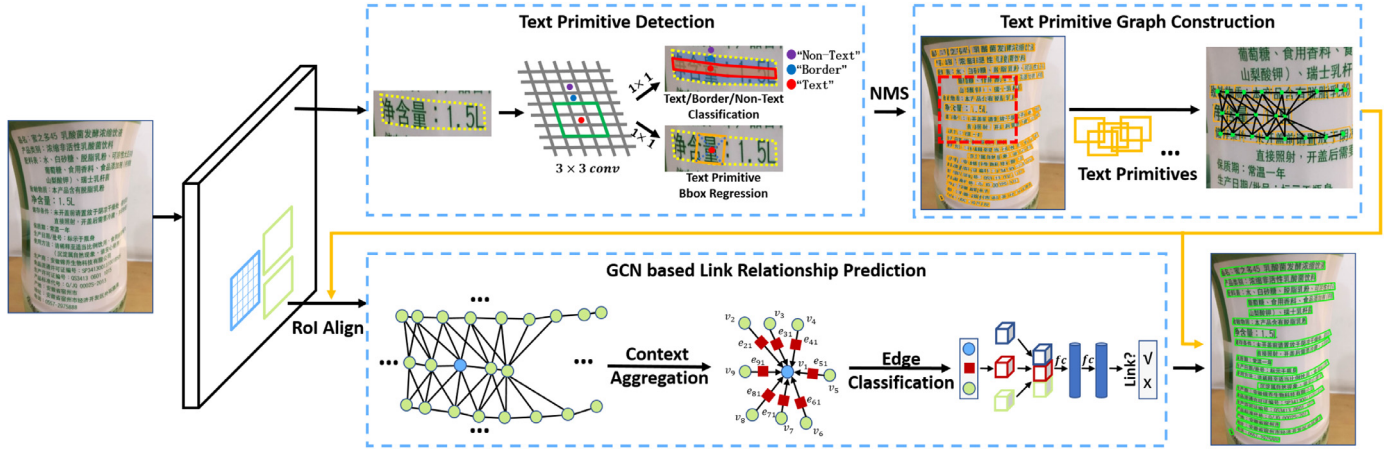


Fig. 2. Overview of our ReLaText. For brevity, the FPN backbone network is omitted and only one of the three parallel forward paths on the feature pyramid levels is illustrated.

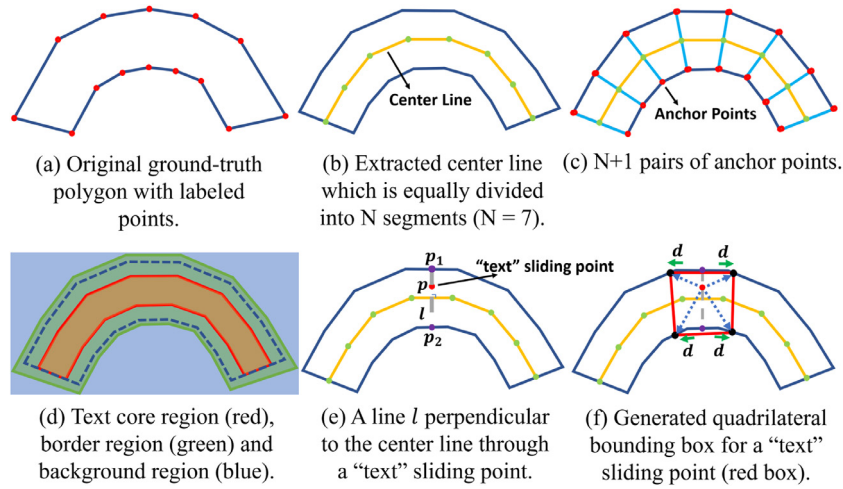


Fig. 3. Illustration of ground-truth bounding box generation for a text primitive. Given a ground-truth polygon, we represent it with $N + 1$ pairs of anchor points following Long et al. [53] as shown in (a)–(c). Then, we shrink and expand the polygon by scaling factors of 0.5 and 1.2 respectively to generate a text core region and a border region (d). Only pixels within the text core region are labeled as “text”. Let p denote a “text” sliding point (red point in (e)) and l denote a line which is perpendicular to the center line through p . Assume that l intersects with two long curved edges of the polygon at points p_1 and p_2 (purple points in (e)). Then, we move p_1 and p_2 forward and backward along the upper and lower curved edges d pixels respectively to get four vertices (black points in (f)) of the corresponding text primitive, where d is set as 12, 24 and 81 for P_2 , P_3 and P_4 , respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

depicted in Fig. 3(e–f) to generate the ground-truth bounding box of its corresponding text primitive. In the inference stage, if a pixel on a pyramid level is classified as “text”, its corresponding detection module will give it a “text” label and directly predict the offsets from it to the vertices of its corresponding quadrilateral text primitive. As depicted in Fig. 2, each detection module is implemented as a 3×3 convolutional layer followed by two sibling 1×1 convolutional layers for text/border/non-text classification and quadrilateral bounding box regression, respectively.

To reduce false alarms, we only keep “text” pixels whose textness scores are higher than a pre-defined threshold which is set as 0.85 in our current implementation. After that, on each pyramid level, we use the standard NMS algorithm with an Intersection-over-Union (IoU) threshold of 0.6 to remove redundant text primitives.

3.3. Text primitive graph construction

We group the detected text primitives from all pyramid levels into a directed graph which is denoted as $\mathbf{G} = \{\mathbf{V}, \mathbf{E}\}$. $\mathbf{V} = \{v_1, v_2, \dots, v_N\}$ is the node set with element v_i denoting the i -th text primitive. $\mathbf{E} = \{e_{i \rightarrow j} = (v_i, v_j) | v_i, v_j \in \mathbf{V}\}$ is the edge set with

element $e_{i \rightarrow j}$ denoting an edge pointing from v_i to v_j . Given N nodes, there are $N(N - 1)$ possible edges between all node pairs. However, if we take all these edges into account, the computation costs for the succeeding link relationship prediction step will be very expensive. Actually, it is unnecessary to predict link relationships between all text primitive pairs as each text instance can be perfectly represented by a sequence of ordered text primitives with link relationships between only nearby text primitive pairs. Moreover, as sizes of text primitives in a same text instance should be similar, we can ignore the link relationships of two text primitives detected from different pyramid levels. According to these assumptions, only pairs of nearby text primitives detected from a same pyramid level are linked with edges. Specifically, given two text primitives v_i and v_j , we define a score to indicate whether there exists a directed edge pointing from v_i to v_j as follows:

$$\text{exist}(e_{i \rightarrow j}) = \begin{cases} 1, & \text{if } v_i \in \text{KNN}(v_j), \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

where $\text{KNN}(v_j)$ denotes the set of K ($K=10$) nearest neighbors of v_j at the same pyramid level. The Euclidean distance between the

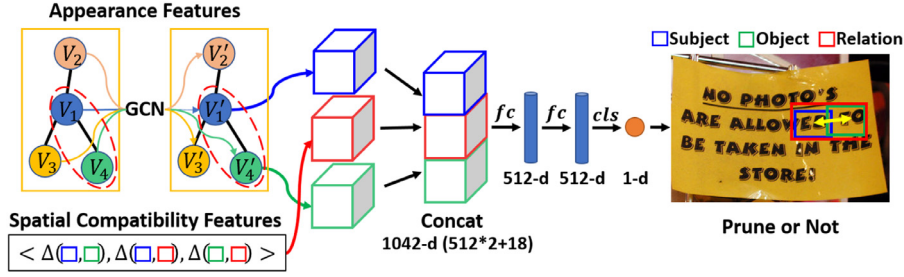


Fig. 4. Architecture of the edge classifier for link relationship prediction.

center points of two text primitives is taken as the distance measure.

In this way, the generated text primitive graph can be divided into three disjoint subgraphs, i.e., \mathbf{G}_{P_2} , \mathbf{G}_{P_3} and \mathbf{G}_{P_4} , which are generated from three feature pyramid levels, P_2 , P_3 and P_4 , respectively. In each subgraph, there are K edges pointing to each node. It should be noted that the neighborhood relationship between a text primitive pair v_i and v_j could be asymmetric, which means it is possible that v_i is one of the K nearest neighbors of v_j but the converse is not true. That is the main reason why the text primitive graph constructed here is a directed graph other than an undirected graph.

3.4. GCN based link relationship prediction

3.4.1. Context-enhanced node representation

For each detected text primitive, we extract a visual feature from its bounding box on its corresponding feature map as its initial node representation. Specifically, for each text primitive v_i , we adopt the RoI Align algorithm [49] to extract a $256 \times 5 \times 5$ feature descriptor from its bounding box on its corresponding pyramid level firstly, which is then fed into a 2-hidden-layer fully connected neural network with 512 nodes at each layer to generate its initial node representation $\mathbf{g}_i^{(0)}$.

After that, we use a graph convolutional neural network to enhance node representations by propagating context between nodes in each text primitive subgraph. Here, we make some modifications to the GCN topology proposed in [40] to improve its capability. Specifically, for a target node v_i in a subgraph \mathbf{G}_{P_k} ($k=2, 3, 4$), the representations of v_i and its neighboring nodes $\mathcal{N}(v_i) = \{v_j | \text{exist}(e_{j \rightarrow i}) = 1\}$ are first transformed via two learned non-linear transformation functions f_v and f_e , respectively. Then, these transformed representations are gathered with predetermined weights α , followed by a non-linear activation function σ . This layer-wise propagation can be written as

$$\mathbf{g}_i^{(l+1)} = \sigma \left(f_v^{(l)}(\mathbf{g}_i^{(l)}) + \sum_{v_j \in \mathcal{N}(v_i)} \alpha_i f_e^{(l)}(\mathbf{g}_j^{(l)}) \right), \quad (2)$$

where $\mathbf{g}_i^{(l)}$ is the updated node representation of v_i output from the l -th GCN layer. $f_v^{(l)}$ and $f_e^{(l)}$ are implemented with multi-layer perceptrons (MLPs) with rectified linear unit (ReLU) between layers. To normalize the transformed neighboring representations, we set α_i as $1/|\mathcal{N}(v_i)|$. And $\sigma(\cdot)$ is implemented as a ReLU activation function.

According to Eq. (2), each node in each graph convolutional layer can only interact with its first-order neighbors. In order to increase the receptive field of each node, we stack L graph convolutional layers to construct an L -layer graph convolutional network. In this way, each node can leverage the information from more distant neighbors in its corresponding subgraph to improve its node representation. The outputs of the last GCN layer are used as the

final node representations, i.e., $\mathbf{g}_i^{(L)}$ is the final node representation of v_i .

3.4.2. Edge classification

The edge classification module is used to prune edges that link two text primitives not belonging to a same text instance. Given an edge $e_{i \rightarrow j}$ directed from v_i to v_j , we extract its feature representation \mathbf{x}_{ij} by concatenating the node representations of v_i and v_j and the spatial compatibility feature of their bounding boxes b_i and b_j , i.e., $\mathbf{x}_{ij} = [\mathbf{g}_i^{(L)}; \mathbf{l}_{ij}; \mathbf{g}_j^{(L)}]$. The spatial compatibility feature \mathbf{l}_{ij} is used to measure the relative scale and location relationships between the bounding boxes of v_i and v_j . Following Zhang et al. [39], let b_{ij} denote the union bounding box of b_i and b_j , then \mathbf{l}_{ij} is defined as an 18-d vector concatenating three 6-d vectors, which indicate the box delta of b_i and b_j , b_i and b_{ij} , b_j and b_{ij} , respectively. Given two bounding boxes $b_i = \{x^i; y^i; w^i; h^i\}$ and $b_j = \{x^j; y^j; w^j; h^j\}$, their box delta is defined as $\Delta(b_i, b_j) = (t_x^{ij}, t_y^{ij}, t_w^{ij}, t_h^{ij}, t_x^{ji}, t_y^{ji}, t_w^{ji}, t_h^{ji})$ where each dimension is given by

$$\begin{aligned} t_x^{ij} &= (x^i - x^j)/w^i, & t_y^{ij} &= (y^i - y^j)/h^i, \\ t_w^{ij} &= \log(w^i/w^j), & t_h^{ij} &= \log(h^i/h^j), \\ t_x^{ji} &= (x^j - x^i)/w^j, & t_y^{ji} &= (y^j - y^i)/h^j. \end{aligned} \quad (3)$$

As depicted in Fig. 4, the edge classifier is a binary classifier which takes the feature representation of each edge as input and outputs whether the input edge should be pruned or not. It is implemented with a 2-hidden-layer MLP with 512 nodes at each hidden layer and 1 Sigmoid activation node at its output layer. As the text primitive graph is a directed graph, there could exist two edges with opposite directions linking a node pair. In our implementation, as long as there exists one edge linking a node pair after edge classification, we consider that there exists a link relationship between this node pair.

To ease implementation, we directly use the ground-truth text primitives to construct the three text primitive subgraphs \mathbf{G}_{P_2} , \mathbf{G}_{P_3} , \mathbf{G}_{P_4} in the training stage. Moreover, we also add some synthesized “non-text” primitives into the subgraphs to improve the robustness of our GCN-based link prediction module to false alarms detected in the inference stage. Specifically, on each FPN pyramid level, we randomly sample a set of “non-text” pixels and assign each of them an artificial bounding box whose width and height are equal to the average width and height of all ground-truth text primitives respectively. Then, we assign all “text” and sampled “non-text” primitives an equal textness score (e.g., 1.0) and randomly shuffle them. After that, we perform a standard NMS algorithm with an IoU threshold of 0.3 on them to generate a node set \mathbf{V}_{P_k} ($k=2, 3, 4$), with which the corresponding subgraph is constructed by using the method proposed in Section 3.3. In each subgraph, we label edges whose two nodes belong to a same text instance as positive, otherwise negative. During training, we ignore all the negative edges that connect two “non-text” nodes, and then adopt the OHEM [80] method to select an equal number of hard

positive and hard negative samples to train our GCN-based link prediction module.

4. Loss function

Multi-task loss for text primitive detection. There are two sibling output layers for each text primitive detection module, i.e., a textness score map prediction layer and a quadrilateral bounding box regression layer. The multi-task loss function is defined as follows:

$$\mathcal{L}_{\text{TPD}_{P_i}} = \frac{1}{N} \sum_j \mathcal{L}_{\text{cls}}(c_j, c_j^*) + \frac{1}{N_{fg}} \sum_k \mathcal{L}_{\text{reg}}(t_k, t_k^*), \quad (4)$$

where N is the number of sampling pixels (including N_{fg} positive ones), c_j and c_j^* are predicted and ground-truth labels for the j -th sampling pixel respectively, $\mathcal{L}_{\text{cls}}(c_j, c_j^*)$ is a binary cross-entropy loss for classification tasks, t_k and t_k^* are predicted and ground-truth 8-d normalized coordinate offsets as stated in [47] for the k -th positive sampling pixel, $\mathcal{L}_{\text{reg}}(t_k, t_k^*)$ is a Smooth-L₁ loss [18] for regression tasks.

The total loss of the text primitive detection module is the sum of the losses of three scale-specific detection modules.

Loss for edge classification. Let E denote the set of selected edges for edge classification training, r_i and r_i^* be the predicted and ground-truth labels for the i -th edge e_i , and $\mathcal{L}(r_i, r_i^*)$ be a binary cross-entropy loss for classification tasks. The loss function for edge classification is defined as follows:

$$\mathcal{L}_{\text{link}} = \frac{1}{|E|} \sum_{e_i \in E} \mathcal{L}(r_i, r_i^*). \quad (5)$$

5. Experiments

5.1. Datasets and evaluation protocols

To evaluate the performance of our proposed ReLaText, we conduct comprehensive experiments on five scene text detection benchmark datasets, including RCTW-17 [81], MSRA-TD500 [82], Total-Text [23], CTW1500 [24] and DAST1500 [30]. We follow the evaluation protocols defined by the authors to make our results comparable to the ones from other methods.

RCTW-17 [81] contains 8,034 training images and 4,229 testing images with scene texts in either Chinese or English. In this dataset, text instances are multi-oriented and labeled by quadrangles in text-line level.

MSRA-TD500 [82] contains 300 training images and 200 testing images, including English and Chinese scripts. Text instances in this dataset are multi-oriented and labeled by rotated rectangles in text-line level. Since the training images are too few, we follow the common practice of previous works to add 400 more images from HUST-TR400 [83] to the training data.

Total-Text [23] contains 1,255 training images and 300 testing images, including 11,459 text instances (4,907 are curved texts) in total. In this dataset, text instances are labeled in word-level with polygons.

CTW1500 [24] contains 1,000 training images and 500 testing images. There are 10,751 text instances in total, where 3,530 are curved texts and at least one curved text per image. In this dataset, text instances are annotated in text-line level with 14-point polygons.

DAST1500 [30] is a dense and arbitrary-shaped text detection dataset, which collects commodity images with detailed descriptions of the commodities on small wrinkled packages from the Internet. It contains 1,038 training images and 500 testing images. Polygon annotations are given in text-line level.

SynthText [11] is a synthetic dataset and consists of about 800k synthetic images generated by blending text instances rendered with random fonts, sizes, colors and orientations in natural images. As the training sets of MSRA-TD500, Total-Text, CTW1500 and DAST1500 are too small, we follow most previous methods [30,53,57] to use this synthetic dataset with word-level annotations to pre-train our text detection models.

5.2. Implementation details

The weights of ResNet-50 related layers in the FPN backbone network are initialized with a pre-trained ResNet-50 model for the ImageNet classification task [79]. The weights of newly added layers are initialized with a Gaussian distribution of mean 0 and standard deviation 0.01. Our ReLaText text detection models are trained in an end-to-end manner and optimized by a standard SGD algorithm with a momentum of 0.9 and weight decay of 0.0005.

The number of training iterations and adjustment strategy of learning rate depend on the sizes of different datasets. Specifically, for RCTW-17, we use the provided 8,034 training images for training and the model is trained for 400k iterations with a base learning rate of 0.004, which is divided by 10 at every 180k iterations. For the other four small scale datasets, we first use the SynthText dataset to pre-train a text detection model for 860k iterations with a base learning rate of 0.004, which is divided by 10 at every 387k iterations. The pre-trained model is then fine-tuned on the corresponding training sets of these four datasets, respectively. All these models are fine-tuned for 70k iterations with a base learning rate of 0.004, which is divided by 10 at every 32k iterations.

We implement ReLaText based on PyTorch³ v0.4.1 and conduct experiments on a workstation with 4 Nvidia V100 GPUs. In each training iteration, we sample one image for each GPU. For each image, we sample a mini-batch of 128 text, 128 border and 128 background pixels for each text primitive detection module. And we select a mini-batch of 64 hard positive and 64 hard negative edges in each subgraph for edge classification training. We adopt a multi-scale training strategy during training. While keeping the aspect ratio, the shorter side of each selected training image is randomly rescaled to a number in {464, 592, 720, 848, 976}. Moreover, when training our text detector on MSRA-TD500 and RCTW-17, we also rotate training images by a random angle in {0°, 90°, 180°, 270°} for data augmentation.

In the testing phase, the detected text primitives are grouped into the whole text instances based on the predicted link relationships with a threshold of 0.7 by using the Union-Find algorithm. Then we use a polynomial curve fitting algorithm to fit the two long sides of a polygon (or a quadrilateral) that encloses all the grouped text primitives for each text instance. Final detection results from different pyramid levels are aggregated by a polygon NMS algorithm [24] with an IoU threshold of 0.2.

5.3. Text detection performance evaluation

In this section, we evaluate our ReLaText on RCTW-17, MSRA-TD500, Total-Text, CTW1500 and DAST1500. In all these experiments, a GCN with 3 layers is applied and all MLPs in the GCN have 2 layers. We compare the performance of our approach with other most competitive results on these five benchmark tasks. For fair comparisons, all the reported results are based on single-model and single-scale testing.

Multilingual and multi-oriented text detection benchmarks. We evaluate our approach on RCTW-17 and MSRA-TD500, which are both multilingual and multi-oriented text detection benchmarks. The longer sides of testing images are set as 1500 and 900

³ <https://pytorch.org/>.

Table 1

Performance comparison on RCTW-17. “L: 1500” means that the longer side of each testing image is resized to be 1500 pixels. * indicates the results are from Liao et al. [85]. †Results listed below represent the accuracy of ReLaText when tested with the image scales in the second column.

Methods	Image scale for testing	RCTW-17			ReLaText†		
		P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
Official baseline [81]	N/A	76.0	40.4	52.8	-	-	-
EAST* [5]	N/A	59.7	47.8	53.1	-	-	-
RRD [85]	N/A	72.4	45.3	55.7	-	-	-
LOMO [29]	L:1024	80.4	50.8	62.3	77.3	57.8	66.1
TextMountain [54]	L:1500	80.8	55.2	65.6	75.9	61.7	68.1
IncepText [84]	N/A	78.5	56.9	66.0	-	-	-
ReLaText (Proposed)	L:1500	75.9	61.7	68.1			

Table 2

Performance comparison on MSRA-TD500. “L: 900” means that the longer side of each testing image is resized to be 900 pixels. †Results listed below represent the accuracy of ReLaText when tested with the image scales in the second column. Note that “FPS” is for reference only because the experimental environments are different.

Methods	Image scale for testing	MSRA-TD500				ReLaText†		
		P(%)	R(%)	F(%)	FPS	P(%)	R(%)	F(%)
EAST [5]	N/A	87.3	67.4	76.1	13.2	-	-	-
SegLink [4]	768 × 768	86.0	70.0	77.0	8.9	90.1	81.6	85.7
Wu et al. [33]	N/A	77.0	78.0	77.0	4.0	-	-	-
PixelLink [50]	768 × 768	83.0	73.2	77.8	3.0	90.1	81.6	85.7
TextSnake [53]	768 × 1280	83.2	73.9	78.3	1.1	86.9	83.0	84.9
TextField [34]	768 × 768	87.4	75.9	81.3	5.2	90.1	81.6	85.7
Lyu et al. [14]	768 × 768	87.6	76.2	81.5	5.7	90.1	81.6	85.7
Tian et al. [56]	L:800	84.2	81.7	82.9	3.0	89.8	81.6	85.5
CRAFT [57]	L:1600	88.2	78.2	82.9	8.6	86.3	79.7	82.9
SBD [46]	1200 × 1600	89.6	80.5	84.8	3.2	86.3	79.7	82.9
ReLaText (Proposed)	L:900	90.5	83.2	86.7	8.3			

Table 3

Performance comparison on Total-Text. “L: 1000” means that the longer side of each testing image is resized to be 1000 pixels, while “S” is for shorter side. †Results listed below represent the accuracy of ReLaText when tested with the image scales in the second column.

Methods	Image scale for testing	Total-Text			ReLaText†		
		P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
CENet [32]	Original Size	59.9	54.4	57.0	78.3	78.8	78.6
Lyu et al. [26]	S:1000	69.0	55.0	61.3	80.4	81.7	81.1
TextSnake [53]	512 × 512	82.7	74.5	78.4	83.3	74.1	78.5
MSR [35]	N/A	83.8	74.8	79.0	-	-	-
TextField [34]	768 × 768	81.2	79.9	80.6	86.4	81.8	84.0
PSENet [51]	L:1280	84.0	78.0	80.9	81.8	82.3	82.0
Mask R-CNN [27]	S:848	81.5	80.5	81.0	81.4	82.0	81.7
SegLink++ [30]	S:768	82.1	80.9	81.5	82.0	81.6	81.8
LOMO [29]	N/A	88.6	75.7	81.6	-	-	-
CRAFT [57]	L:1280	87.6	79.9	83.6	81.8	82.3	82.0
ReLaText (Proposed)	L:1000	84.8	83.1	84.0			

for RCTW-17 and MSRA-TD500, respectively. As shown in Table 1 and Table 2, ReLaText outperforms the closest methods [84] and [46] substantially by improving F-score from 66.0% to 68.1% and 84.8% to 86.7% on RCTW-17 and MSRA-TD500, respectively. Moreover, as the scales of testing images affect the accuracy of text detection models significantly, we also report the results of our text detector when tested with the same image scales as previous methods. The results are presented in the last three columns of Table 1 and Table 2, from which we can find that, even without model retraining, our text detector can still achieve better results than most of its counterparts when tested with the same image scales.

Curved text detection benchmarks. We conduct experiments on two curved-text focused datasets, namely Total-Text and CTW1500, to evaluate the performance of ReLaText on detecting arbitrary-shaped texts in natural scene images. The longer sides

of testing images are set as 1000 and 800 for Total-Text and CTW1500, respectively. Note that two latest methods [27,28] adopt more powerful backbone networks, i.e., supervised pyramid context network in [27] and pyramid attention network in [28] to push the text detection performance of Mask R-CNN on Total-Text and CTW1500, respectively. For fairer comparisons, we compare our approach with these two Mask R-CNN based methods by using the same ResNet50-FPN backbone network. As shown in Table 3 and Table 4, ReLaText achieves the best F-score of 84.0% on Total-Text and 84.8% on CTW1500. Moreover, when using the same testing scales as other methods, ReLaText can still achieve better or comparable results.

Dense and arbitrary-shaped text detection benchmark. We further evaluate the performance of ReLaText on the newly proposed DAST1500 dataset, which contains a large number of dense and arbitrary-shaped text instances. Following other methods on

Table 4

Performance comparison on CTW1500. “L: 800” means that the longer side of each testing image is resized to be 800 pixels, while “S” is for shorter side. * indicates that the result is provided by author(s). †Results listed below represent the accuracy of ReLaText when tested with the image scales in the second column.

Methods	Image scale for testing	CTW1500				ReLaText†		
		P(%)	R(%)	F(%)	FPS	P(%)	R(%)	F(%)
CTD+TLOC [24]	600 × 1000	77.4	69.8	73.4	13.3	86.3	83.3	84.8
TextSnake [53]	Original Size	67.9	85.3	75.6	-	85.9	76.2	80.8
LOMO [29]	L:512	89.2	69.6	78.4	4.4	87.6	78.2	82.6
Tian et al. [56]	L:800	82.7	77.8	80.1	-	86.2	83.3	84.8
Wang et al. [25]	720 × 1280	80.1	80.2	80.1	10.0	84.7	83.6	84.2
SegLink++ [30]	S:512	82.8	79.8	81.3	-	86.4	82.7	84.5
TextField [34]	576 × 576	83.0	79.8	81.4	-	87.6	80.3	83.8
MSR [35]	N/A	85.0	78.3	81.5	-	-	-	-
PSENet [51]	L:1280	84.8	79.7	82.2	3.9	84.1	83.4	83.8
Mask R-CNN* [28]	S:512	83.8	81.7	82.7	-	86.4	82.7	84.5
CRAFT [57]	L:1024	86.0	81.1	83.5	-	84.8	82.8	83.8
ReLaText (Proposed)	L:800	86.2	83.3	84.8	10.6			



Fig. 5. Qualitative detection results of ReLaText. (a-b) are from RCTW-17, (c-d) are from MSRA-TD500, (e-f) are from Total-Text, (g-h) are from CTW1500, and (i-j) are from DAST1500.

Table 5

Performance comparison on DAST1500. * indicates the results are from [30].

Methods	P(%)	R(%)	F(%)
SegLink* [4]	66.0	64.7	65.3
CTD+TLOC* [24]	73.8	60.8	66.6
PixelLink* [50]	74.5	75.0	74.7
SegLink++ [30]	79.6	79.2	79.4
ReLaText (Proposed)	89.0	82.9	85.8

this dataset, we resize the testing images to 768×768 . As shown in Table 5, ReLaText achieves the best result of 89.0%, 82.9% and 85.8% in precision, recall and F-score, respectively, outperforming other competing approaches significantly.

Inference time. Based on our current implementation, ReLaText has an inference time of 0.29s, 0.12s, 0.31s, 0.09s and 0.23s per image when using a single V100 GPU for $L=1500$, $L=900$, $L=1000$, $L=800$ and $L=768$ on RCTW-17, MSRA-TD500, Total-Text, CTW1500 and DAST1500, respectively.

Qualitative results. The superior performance achieved on the above five datasets demonstrates the effectiveness and robustness of ReLaText. As shown in Fig. 5, ReLaText can detect scene texts under various challenging conditions, such as non-uniform illumi-

nation, low contrast, low resolution, large inter-character spacing, small inter-line spacing and arbitrary shapes.

5.4. Ablation study

Influence of the network depth. We investigate the influences of the depths of GCN and MLPs in GCN layers to text detection accuracies, respectively. The experimental results on CTW1500 are shown in Table 6. Since less parameters harm the performance and more parameters do not bring extra gains, setting the depth of GCN as 3 and the depth of each MLP as 2 exhibits a good trade-off between performance and efficiency for our approach.

Influence of the number of neighbors. We also investigate the influence of the number of neighbors for graph construction. The experimental results on CTW1500 are shown in Table 7. When the number of neighbors is in the range of 5 to 30, our model can achieve reasonable and comparable performance. However, a too small number (≤ 5), which means an aggressive early-pruning, will lead to a significant performance degradation.

Why use the ground-truth text primitives to train GCN? As stated in Section 3.4.2, we directly use the ground-truth text primitives to generate training samples for GCN. However, it is also possible to use the detected text primitives from the previous module to generate training samples. So we have also conducted ablation experiments on the RCTW-17 dataset to compare these two meth-

Table 6
Ablation study of the depth of GCN and the depth of each MLP in GCN.

Depth of GCN	P(%)	R(%)	F(%)	Depth of MLP	P(%)	R(%)	F(%)
1	85.5	82.5	84.0	1	85.0	82.8	83.9
2	86.2	82.4	84.2	2	86.2	83.3	84.8
3	86.2	83.3	84.8	3	86.0	82.9	84.4
4	86.5	82.5	84.5	4	86.0	82.5	84.2

Table 7
Ablation study of the number of neighbors for graph construction.

# of neighbors	1	2	5	10	20	30
F(%)	19.8	80.8	84.0	84.8	84.7	84.3

Table 8
Comparison of two training sample generation methods for training GCN.

Source of Training Samples	P(%)	R(%)	F(%)	Training Time
Ground-truth (ReLaText)	75.9	61.7	68.1	14.0h
Detected	75.5	61.2	67.6	18.5h

ods. According to the experimental results in Table 8, we find that using ground-truth text primitives to train our GCN model is a better solution as it could achieve slightly higher text detection accuracy and improve the training speed significantly.

Effectiveness of visual relationship prediction based text-line grouping. We compare our two visual relationship prediction based text-line grouping methods with two previously widely used text-line grouping methods, which are based on 8-neighborhood pixel connectivity and linkage respectively, on CTW1500 and RCTW-17 to demonstrate the effectiveness of the visual relationship prediction based problem formulation for text detection. In our conference paper [38], we used a relation network to do link relationship prediction. In this work, we propose to exploit a GCN model to further improve the accuracy of link relationship prediction. For a fair comparison, we replace our line grouping module with the 8-neighborhood pixel connectivity and linkage based

methods in our codes and tune the hyper-parameters carefully to get their best possible results. The quantitative results are given in Table 9, from which we can observe that the performance of pixel connectivity and linkage based methods are obviously inferior to our link relationship based methods, i.e., 82.6% and 82.7% vs. 83.8% and 84.8% on CTW1500, 66.2% and 66.2% vs. 67.0% and 68.1% on RCTW-17, which demonstrates the superiority of visual relationship prediction based text-line grouping.

Effectiveness of GCN-based link relationship prediction. We compare the GCN-based link relationship prediction with our previous relation network based link relationship prediction to demonstrate the effectiveness of GCN for link relationship prediction. As shown in the last two rows in Table 9, exploiting GCN to predict link relationship is more effective than relation network, and the F-score is improved substantially from 83.8% to 84.8% on CTW1500 and from 67.0% to 68.1% on RCTW-17 respectively. Some comparison examples are shown in Fig. 6. Based on our observations, ReLaText can leverage context information more effectively to improve link relationship prediction accuracy, leading to improved robustness to text instances with large inter-character or small inter-line spacing.

5.5. Limitations of our approach

Although our ReLaText shows superior capability in most scenarios as demonstrated in the previous experiments, it still fails in some difficult cases such as text-lines with ambiguous layouts and extremely small texts. Some failure examples are presented in Fig. 7. Note that these difficulties are common challenges for other state-of-the-art methods. Finding effective solutions to these problems will be our future work.

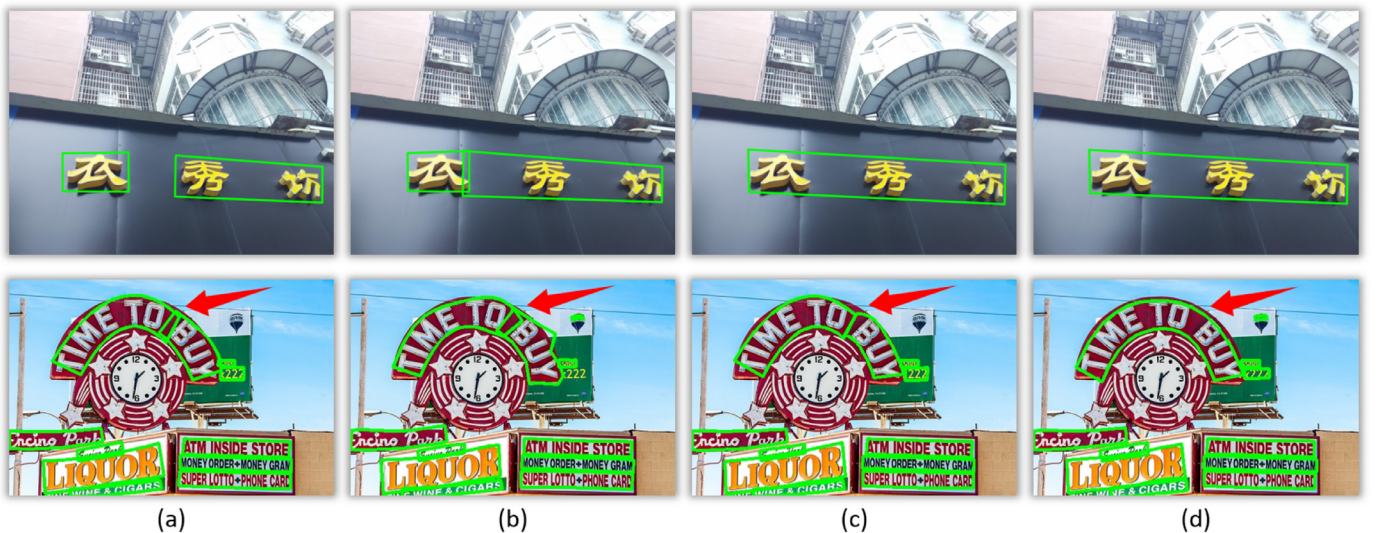


Fig. 6. Some comparison examples from different text-line grouping approaches: (a) Connectivity; (b) Linkage; (c) RN based link relationship prediction; (d) GCN based link relationship prediction (ReLaText).

Table 9

Comparison of different text-line grouping strategies. “Connectivity”, “Linkage”, “RN based link relationship”, “GCN based link relationship” denote grouping based on 8-neighborhood pixel connectivity, 8-neighborhood pixel linkage, link relationships predicted by Relation Network and GCN, respectively.

Methods	CTW1500			RCTW-17		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
Connectivity (e.g., [33,35,53])	83.6	81.6	82.6	76.0	58.6	66.2
Linkage (e.g., [4,30,50])	84.0	81.4	82.7	76.8	58.1	66.2
RN based link relationship [38]	85.0	82.5	83.8	76.4	59.6	67.0
GCN based link relationship (ReLaText)	86.2	83.3	84.8	75.9	61.7	68.1



Fig. 7. Some failure examples. Ground-truths are depicted in red, and detection results are depicted in green. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

6. Conclusion and future work

In this paper, we introduce a new arbitrary-shaped text detection approach by formulating text detection as a visual relationship detection problem and using a GCN-based visual relationship detection framework to solve the challenging text-line grouping problem. As GCN can effectively leverage context information to improve link prediction accuracy, our visual relationship prediction based text-line grouping approach can achieve better text detection accuracy than previous local pixel connectivity based methods, especially when dealing with text instances with large inter-character or very small inter-line spacing. Consequently, the proposed ReLaText has achieved state-of-the-art performance on five public benchmark datasets, namely RCTW-17, MSRA-TD500, Total-Text, CTW1500 and DAST1500.

For future work, we will explore other kinds of relationships, such as a “duplicate” relationship to suppress duplicate text instances, to improve text detection accuracy. We will also study how to integrate our text detector into other visual relationship detection based document understanding systems to help improve their accuracies. Furthermore, we will develop better text primitive detection methods to improve the robustness of our text detector to extremely small or extremely large text instances.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at [10.1016/j.patcog.2020.107684](https://doi.org/10.1016/j.patcog.2020.107684).

References

- [1] W. He, X. Zhang, F. Yin, C. Liu, Deep direct regression for multi-oriented scene text detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 745–753.
- [2] M. Liao, B. Shi, X. Bai, X. Wang, W. Liu, Textboxes: A fast text detector with a single deep neural network, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2017, pp. 4161–4167.
- [3] Y. Liu, L. Jin, Deep matching prior network: Toward tighter multi-oriented text detection, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 1962–1969.
- [4] B. Shi, X. Bai, S. Belongie, Detecting oriented text in natural images by linking segments, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 2550–2558.
- [5] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, J. Liang, East: An efficient and accurate scene text detector, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 5551–5560.
- [6] L. Neumann, J. Matas, A method for text localization and recognition in real-world images, in: Asian Conference on Computer Vision, 2010, pp. 770–783.
- [7] X. Yin, X. Yin, K. Huang, H. Hao, Robust text detection in natural scene images, IEEE Trans Pattern Anal Mach Intell 36 (5) (2013) 970–983.
- [8] B. Epshtein, E. Ofek, Y. Wexler, Detecting text in natural scenes with stroke width transform, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2010, pp. 2963–2970.
- [9] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, X. Bai, Multi-oriented text detection with fully convolutional networks, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016, pp. 4159–4167.
- [10] M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman, Reading text in the wild with convolutional neural networks, Int J Comput Vis 116 (1) (2016) 1–20.
- [11] A. Gupta, A. Vedaldi, A. Zisserman, Synthetic data for text localisation in natural images, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016, pp. 2315–2324.
- [12] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, X. Xue, Arbitrary-oriented scene text detection via rotation proposals, IEEE Trans Multimedia 20 (11) (2018) 3111–3122.
- [13] Z. Tian, W. Huang, T. He, P. He, Y. Qiao, Detecting text in natural image with connectionist text proposal network, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 56–72.
- [14] P. Lyu, C. Yao, W. Wu, S. Yan, X. Bai, Multi-oriented scene text detection via corner localization and region segmentation, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 7553–7563.
- [15] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, J. Yan, Fots: Fast oriented text spotting with a unified network, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 5676–5685.

- [16] Z. Zhong, L. Sun, Q. Huo, Improved localization accuracy by locnet for faster r-cnn based text detection in natural scene images, *Pattern Recognit* 96 (2019) 106986.
- [17] W. He, X. Zhang, F. Yin, Z. Luo, J.-M. Ogier, C. Liu, Realtime multi-scale scene text detection with scale-based region proposal network, *Pattern Recognit* 98 (2020) 107026.
- [18] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.
- [19] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: *Proceedings of the European Conference on Computer Vision*, 2016, pp. 21–37.
- [20] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.
- [21] L. Huang, Y. Yang, Y. Deng, Y. Yu, Densebox: unifying landmark localization with end to end object detection, *arXiv preprint arXiv:1509.04874* (2015).
- [22] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.
- [23] C.K. Ch'ng, C.S. Chan, Total-text: A comprehensive dataset for scene text detection and recognition, in: *Proceedings of the International Conference on Document Analysis and Recognition*, 2017, pp. 935–942.
- [24] Y. Liu, L. Jin, S. Zhang, C. Luo, S. Zhang, Curved scene text detection via transverse and longitudinal sequence connection, *Pattern Recognit* 90 (2019) 337–345.
- [25] X. Wang, Y. Jiang, Z. Luo, C. Liu, H. Choi, S. Kim, Arbitrary shape scene text detection with adaptive text region representation, in: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2019, pp. 6449–6458.
- [26] P. Lyu, M. Liao, C. Yao, W. Wu, X. Bai, Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes, in: *Proceedings of the European Conference on Computer Vision*, 2018, pp. 67–83.
- [27] E. Xie, Y. Zang, S. Shao, G. Yu, C. Yao, G. Li, Scene text detection with supervised pyramid context network, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 9038–9045.
- [28] Z. Huang, Z. Zhong, L. Sun, Q. Huo, Mask r-cnn with pyramid attention network for scene text detection, in: *Proceedings of the IEEE Winter conference on Applications of Computer Vision*, 2019, pp. 764–772.
- [29] C. Zhang, B. Liang, Z. Huang, M. En, J. Han, E. Ding, X. Ding, Look more than once: An accurate detector for text of arbitrary shapes, in: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2019, pp. 10552–10561.
- [30] J. Tang, Z. Yang, Y. Wang, Q. Zheng, Y. Xu, X. Bai, Seglink++: detecting dense and arbitrary-shaped scene text by instance-aware component grouping, *Pattern Recognit* 96 (2019) 106954.
- [31] H. Hu, C. Zhang, Y. Luo, Y. Wang, J. Han, E. Ding, Wordsup: Exploiting word annotations for character based text detection, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4940–4949.
- [32] J. Liu, C. Zhang, Y. Sun, J. Han, E. Ding, Detecting text in the wild with deep character embedding network, in: *Asian Conference on Computer Vision*, 2018, pp. 501–517.
- [33] Y. Wu, P. Natarajan, Self-organized text detection with minimal post-processing via border learning, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5000–5009.
- [34] Y. Xu, Y. Wang, W. Zhou, Y. Wang, Z. Yang, X. Bai, Textfield: learning a deep direction field for irregular scene text detection, *IEEE Trans. Image Process.* 28 (11) (2019) 5566–5579.
- [35] C. Xue, S. Lu, W. Zhang, Msr: Multi-scale shape regression for scene text detection, in: *Proceedings of the International Joint Conference on Artificial Intelligence*, 2019, pp. 20–36.
- [36] B. Davis, B. Morse, S. Cohen, B. Price, C. Tensmeyer, Deep visual template-free form parsing, in: *Proceedings of the International Conference on Document Analysis and Recognition*, 2019, pp. 134–141.
- [37] S.R. Qasim, H. Mahmood, F. Shafait, Rethinking table recognition using graph neural networks, in: *Proceedings of the International Conference on Document Analysis and Recognition*, 2019, pp. 142–147.
- [38] C. Ma, Z. Zhong, L. Sun, Q. Huo, A relation network based approach to curved text detection, in: *Proceedings of the International Conference on Document Analysis and Recognition*, 2019, pp. 707–713.
- [39] J. Zhang, M. Elhoseiny, S. Cohen, W. Chang, A. Elgammal, Relationship proposal networks, in: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 5678–5686.
- [40] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: *International Conference on Learning Representations*, 2017.
- [41] P. Shivakumara, H.T. Basavaraju, D. Guru, C.L. Tan, Detection of curved text in video: Quad tree based method, in: *Proceedings of the International Conference on Document Analysis and Recognition*, 2013, pp. 594–598.
- [42] J. Fabrizio, M. Robert-Seidowsky, S. Dubuisson, S. Calarasanu, R. Boissel, Textcatcher: a method to detect curved and challenging text in natural scenes, *Int. J. Doc. Anal. Recogn.* 19 (2) (2016) 99–117.
- [43] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.
- [44] L. Gómez, D. Karatzas, Textproposals: a text-specific selective search algorithm for word spotting in the wild, *Pattern Recognit* 70 (2017) 60–74.
- [45] Z. Zhong, L. Jin, S. Huang, Deeptext: A new approach for text proposal generation and text detection in natural images, in: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 1208–1212.
- [46] Y. Liu, S. Zhang, L. Jin, L. Xie, Y. Wu, Z. Wang, Omnidirectional scene text detection with sequential-free box discretization, in: *Proceedings of the International Joint Conference on Artificial Intelligence*, 2019, pp. 3052–3058.
- [47] Z. Zhong, L. Sun, Q. Huo, An anchor-free region proposal network for faster R-CNN-based text detection approaches, *Int. J. Doc. Anal. Recogn.* 22 (3) (2019) 315–327.
- [48] J. Dai, Y. Li, K. He, J. Sun, R-FCN: Object detection via region-based fully convolutional networks, in: *Advances in Neural Information Processing Systems*, 2016, pp. 379–387.
- [49] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.
- [50] D. Deng, H. Liu, X. Li, D. Cai, Pixellink: Detecting scene text via instance segmentation, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, pp. 6773–6780.
- [51] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, S. Shao, Shape robust text detection with progressive scale expansion network, in: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2019, pp. 9336–9345.
- [52] Z. Liu, G. Lin, S. Yang, F. Liu, W. Lin, W.L. Goh, Towards robust curve text detection with conditional spatial expansion, in: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2019, pp. 7269–7278.
- [53] S. Long, J. Ruan, W. Zhang, X. He, W. Wu, C. Yao, Textsnake: A flexible representation for detecting text of arbitrary shapes, in: *Proceedings of the European Conference on Computer Vision*, 2018, pp. 20–36.
- [54] Y. Zhu, J. Du, Textmountain: accurate scene text detection via instance segmentation, *Pattern Recognit* (2020) 107336.
- [55] Z. Liu, G. Lin, S. Yang, J. Feng, W. Lin, W.L. Goh, Learning markov clustering networks for scene text detection, in: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6936–6944.
- [56] Z. Tian, M. Shu, P. Lyu, R. Li, C. Zhou, X. Shen, J. Jia, Learning shape-aware embedding for scene text detection, in: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2019, pp. 4234–4243.
- [57] Y. Baek, B. Lee, D. Han, S. Yun, H. Lee, Character region awareness for text detection, in: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2019, pp. 9365–9374.
- [58] C. Lu, R. Krishna, M. Bernstein, L. Fei-Fei, Visual relationship detection with language priors, in: *Proceedings of the European Conference on Computer Vision*, 2016, pp. 852–869.
- [59] M. Mahdavi, M. Condon, K. Davila, R. Zanibbi, Lpqa: Line-of-sight parsing with graph-based attention for math formula recognition, in: *Proceedings of the International Conference on Document Analysis and Recognition*, 2019, pp. 647–654.
- [60] Y. Li, W. Ouyang, X. Wang, X. Tang, Vip-cnn: Visual phrase guided convolutional neural network, in: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1347–1356.
- [61] Y. Zhu, S. Jiang, Deep structured learning for visual relationship detection, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, pp. 7623–7630.
- [62] R. Yu, A. Li, V.I. Morariu, L.S. Davis, Visual relationship detection with internal and external linguistic knowledge distillation, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1974–1982.
- [63] D. Xu, Y. Zhu, C.B. Choy, L. Fei-Fei, Scene graph generation by iterative message passing, in: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 5410–5419.
- [64] J. Yang, J. Lu, S. Lee, D. Batra, D. Parikh, Graph r-cnn for scene graph generation, in: *Proceedings of the European Conference on Computer Vision*, 2018, pp. 670–685.
- [65] B. Dai, Y. Zhang, D. Lin, Detecting visual relationships with deep relational networks, in: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 3076–3086.
- [66] S. Woo, D. Kim, D. Cho, I.S. Kweon, Linknet: Relational embedding for scene graph, in: *Advances in Neural Information Processing Systems*, 2018, pp. 560–570.
- [67] B. Zhuang, L. Liu, C. Shen, I. Reid, Towards context-aware interaction recognition for visual relationship detection, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 589–598.
- [68] P. Goyal, E. Ferrara, Graph embedding techniques, applications, and performance: a survey, *Knowl Based Syst* 151 (2018) 78–94.
- [69] A. Sperduti, A. Starita, Supervised neural networks for the classification of structures, *IEEE Trans. Neural Networks* 8 (3) (1997) 714–735.
- [70] J. Bruna, W. Zaremba, A. Szlam, Y. Lecun, Spectral networks and locally connected networks on graphs, in: *International Conference on Learning Representations*, 2014.
- [71] M. Henaff, J. Bruna, Y. LeCun, Deep convolutional networks on graph-structured data, *arXiv preprint arXiv:1506.05163* (2015).
- [72] J. Atwood, D. Towsley, Diffusion-convolutional neural networks, in: *Advances in Neural Information Processing Systems*, 2016, pp. 1993–2001.
- [73] M. Niepert, M. Ahmed, K. Kutzkov, Learning convolutional neural networks for graphs, in: *International Conference on Machine Learning*, 2016, pp. 2014–2023.
- [74] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, in: *Advances in Neural Information Processing Systems*, 2017, pp. 1024–1034.

- [75] D.K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R.P. Adams, Convolutional networks on graphs for learning molecular fingerprints, in: *Advances in Neural Information Processing Systems*, 2015, pp. 2224–2232.
- [76] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, in: *International Conference on Learning Representations*, 2018.
- [77] Y. Wang, Y. Sun, Z. Liu, S.E. Sarma, M.M. Bronstein, J.M. Solomon, Dynamic graph cnn for learning on point clouds, *ACM Trans Graph* 38 (5) (2019) 1–12.
- [78] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.
- [79] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [80] A. Shrivastava, A. Gupta, R. Girshick, Training region-based object detectors with online hard example mining, in: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 761–769.
- [81] B. Shi, C. Yao, M. Liao, M. Yang, P. Xu, L. Cui, S. Belongie, S. Lu, X. Bai, Icdar2017 competition on reading chinese text in the wild (rctw-17), in: *Proceedings of the International Conference on Document Analysis and Recognition*, 2017, pp. 1429–1434.
- [82] C. Yao, X. Bai, W. Liu, Y. Ma, Z. Tu, Detecting texts of arbitrary orientations in natural images, in: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2012, pp. 1083–1090.
- [83] C. Yao, X. Bai, W. Liu, A unified framework for multioriented text detection and recognition, *IEEE Trans. Image Process.* 23 (11) (2014) 4737–4749.
- [84] Q. Yang, M. Cheng, W. Zhou, Y. Chen, M. Qiu, W. Lin, W. Chu, Inceptext: A new inception-text module with deformable psroi pooling for multi-oriented scene text detection, in: *Proceedings of the International Joint Conference on Artificial Intelligence*, 2018, pp. 1071–1077.
- [85] M. Liao, Z. Zhu, B. Shi, G. Xia, X. Bai, Rotation-sensitive regression for oriented scene text detection, in: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 5909–5918.

Chixiang Ma is a Ph.D. candidate of University of Science and Technology of China (USTC), Hefei, China. He received the B.Eng. degree in electronic information engineering from USTC in 2017. His research interests include machine learning, pattern recognition, text detection, table detection and recognition.

Lei Sun is a Principal Researcher of the Speech Group in Microsoft Research Asia (MSRA), Beijing, China. He received the B.Eng. degree from University of Science and Technology of China (USTC), Hefei, China, in 2009 and the Ph.D. degree from USTC in 2015. His research interests mainly lie in text detection, layout analysis, table detection and recognition, OCR, document image understanding and information retrieval.

Zhuoyao Zhong is a Researcher of the Speech Group in Microsoft Research Asia (MSRA), Beijing, China. He received the B.Eng. degree in Information Engineering from South China University of Technology (SCUT), Guangzhou, China, in 2015 and the Ph.D. degree in Information and Communication Engineering from SCUT in 2020. His research interests mainly include machine learning, pattern recognition, character recognition, text detection and recognition, layout analysis and text-centric image understanding.

Qiang Huo is the Principal Research Manager of the Speech Group in Microsoft Research Asia (MSRA), Beijing, China. Prior to joining MSRA in August 2007, he had been a faculty member at the Department of Computer Science, The University of Hong Kong since 1998, where he also did his Ph.D. research on speech recognition during 1991 to 1994. From 1995 to 1997, Dr. Huo worked at Advanced Telecommunications Research Institute (ATR), Kyoto, Japan. Over the past 30 years, he has been active in research and making contributions in the fields of speech recognition, handwriting recognition, OCR, gesture recognition, biometric-based user authentication, hardware design for speech and image processing. Dr. Huo received a B.Eng. degree from University of Science and Technology of China (USTC), Hefei, China, in 1987, an M.Eng. degree from Zhejiang University, Hangzhou, China, in 1989, and a Ph.D. degree from USTC in 1994, all in electrical engineering.